AGARD-CP-505

AGARD-CP-505

AGARD

ADVISORY GROUP FOR AEROSPACE RESEARCH & DEVELOPMENT

7 RUE ANCELLE 92200 NEUILLY SUR SEINE FRANCE

AGARD CONFERENCE PROCEEDINGS 505

Bringing down the Barriers to Information Transfer

(L'Abaissement des Barrières s'Opposant au Transfert de l'Information)

Papers presented at the Technical Information Panel Specialists' Meeting held at the Instituto Nacional de Industria, Madrid, Spain 8th–9th October 1991.

NORTH ATLANTIC TREATY ORGANIZATION

Published February 1992

Distribution and Availability on Back Cover

AGARD-CP-505

AGARD

ADVISORY GROUP FOR AEROSPACE RESEARCH & DEVELOPMENT

7 RUE ANCELLE 92200 NEUILLY SUR SEINE FRANCE

AGARD CONFERENCE PROCEEDINGS 505

Bringing down the Barriers to Information Transfer

(L'Abaissement des Barrières s'Opposant au Transfert de l'Information)

Papers presented at the Technical Information Panel Specialists' Meeting held at the Instituto Nacional de Industria, Madrid, Spain 8th—9th October 1991.

North Atlantic Treaty Organization

Organisation du Traité de l'Atlantique Nord

The Mission of AGARD

According to its Charter, the mission of AGARD is to bring together the leading personalities of the NATO nations in the fields of science and technology relating to aerospace for the following purposes:

- Recommending effective ways for the member nations to use their research and development capabilities for the common benefit of the NATO community;
- Providing scientific and technical advice and assistance to the Military Committee in the field of aerospace research and development (with particular regard to its military application);
- Continuously stimulating advances in the aerospace sciences relevant to strengthening the common defence posture;
- Improving the co-operation among member nations in aerospace research and development;
- Exchange of scientific and technical information;
- Providing assistance to member nations for the purpose of increasing their scientific and technical potential;
- Rendering scientific and technical assistance, as requested, to other NATO bodies and to member nations in connection with research and development problems in the aerospace field.

The highest authority within AGARD is the National Delegates Board consisting of officially appointed senior representatives from each member nation. The mission of AGARD is carried out through the Panels which are composed of experts appointed by the National Delegates, the Consultant and Exchange Programme and the Aerospace Applications Studies Programme. The results of AGARD work are reported to the member nations and the NATO Authorities through the AGARD series of publications of which this is one.

Participation in AGARD activities is by invitation only and is normally limited to citizens of the NATO nations.

The content of this publication has been reproduced
directly from material supplied by AGARD or the authors.

Published February 1992

ISBN 92-835-0655-3

Printed by Specialised Printing Services Limited
40 Chigwell Lane, Loughton, Essex IG10 3TZ

Theme

The last few years have seen rapid advances in a number of new technologies aimed at improving the storage and retrieval of information. Information Scientists may be aware of their potential but may have reservations concerning the absence of suitable standards or the short life associated with some developments.

This Specialists' Meeting brought together a group of experts to explain the practicalities of applying these new, powerful technologies, their successes and failures. Topics addressed included Artificial Intelligence, CD ROM, Hypertext and Local Area Networks, within the framework of finding practical solutions to the problems of achieving efficient and effective information transfer. The meeting had the following objectives:

1) To present first-hand, practical experience of initiatives to improve Information Transfer, by experts in the field; And

2) To provide up-to-date information on the new technologies which have been successfully used to promote efficient and effective information transfer.

The meeting was directed particularly at information providers and users, especially those acting within the NATO Community.

Thème

Ces dernières années ont été marquées par l'essor d'un certain nombre de nouvelles technologies visant à améliorer le stockage et la recherche d'informations. Bien que les documentalistes soient conscients des possibilités de ces technologies, ils émettent parfois des réserves sur l'absence de normes adéquates dans ce domaine et sur la durée de vie relativement courte de certains de ces développements.

Cette réunion de spécialistes a rassemblé un groupe d'cxperts chargé d'examiner les détails pratiques de la mise en oeuvre de ces nouvelles technologies performantes et de dresser le bilan des succès et des échecs. Les sujets abordés ont été l'Intelligence artificielle, l'Hypertext, le CD ROM et les réseaux locaux (LAN). Ces sujets ont été examinés dans l'optique d'une recherche de solutions pratiques aux problèmes de la transmission effective et efficace des données. La réunion a eu pour objectif:

- de présenter des enseignements pratiques tirés d'initiatives prises par des spécialistes dans le domaine de la transmission des données
- de fournir des informations actualisées sur les nouvelles technologies qui ont été employées avec succès pour la transmission effective et efficace des données.

La réunion s'est adressée donc, en particulier, aux fournisseurs et aux utilisateurs de l'information, dans le contexte des échanges entre pays membres de l'OTAN.

Technical Information Panel

Chairman: Mr Albert Yanez
Conseiller du Directeur
C.E.D.O.C.A.R
00460 Armées
France

Deputy Chairman: Mr Richard Searle
Chief Librarian
Royal Aerospace Establishment
Farnborough
Hants GU14 6TD
United Kingdom

MEETING PLANNING COMMITTEE

Chairman:	Mr A. del Rey	(SP)
Deputy Chairman:	Mr M.R.C. Wilkinson	(UK)
Members:	Mr M. Schryer	(CA)
	Gen. F. Chevalier	(FR)
	Mr R. Bernhardt	(GE)
	Ms A.M. Correia	(PO)
	Ms D. Deleuze	(SP)
	Mrs M.C. Gutierrez	(SP)
	Mr C. Bigger	(UK)
	Ms G. Cotter	(US)

PANEL EXECUTIVE

Mr G.W. Hart

Mail from Europe:
AGARD–OTAN
Attn: TIP Executive
7, rue Ancelle
92200 Neuilly-sur-Seine
France

Mail from US and Canada:
AGARD–NATO/TIP
Unit 21551
APO AE 09777

Tel: 33 (1) 47 38 57 95
Telex: 610176 (France)
Telefax: 33 (1) 47 38 57 99

Contents

	Page
Theme/Thème	iii
Technical Information Panel	iv
	Reference
Technical Evaluation Report by W. Blados	T
Keynote Address by A. del Rey	K
Local Area Networks (LAN) by G. Farquhar	1
Gateways "Intelligents" par C. Fluhr et C. Machard	2
Utilisation des Systèmes Experts dans des Interfaces pour la Recherche Documentaire (Use of Expert Systems as User Interfaces in Information Retrieval) par J-C. Bassano, D. Archambault, G. Desrocques et A. Mekaouche	3
Search Strategies in Natural Language by B.H.A. Zijlstra	4
Non Boolean Search Methods in Information Retrieval by C.J. van Rijsbergen	5
Hypertext and Hypermedia Systems in Information Retrieval by K.M. Kaye and A.D. Kuhn	6
Automated Input into Databases: OCR and Descriptive Cataloguing by G.E. Knorz	7
Data Compression Techniques by R.A. Hogendoorn	8
Computerized Property Data for Engineering Materials — An Overview by G.H. Wood	9
Facilitating the Transfer of Scientific and Technical Information with Scientific and Technical Numeric Databases by H. Haller	10
List of Participants	A

TECHNICAL EVALUATION REPORT

Walter Blados
Scientific and Technical Information Program
NASA Headquarters (Code JTT)
Washington DC 20546
USA

SUMMARY

This Technical Evaluation Report is in two sections. Section 1 provides a brief summary of each of the papers as presented. Section 2 comprises comments, conclusions, and recommendations partly arising as a consequence of the meeting.

SECTION 1

SUMMARY OF PAPERS AS PRESENTED

KEYNOTE ADDRESS

The author provided a brief overview of the topics that were to be presented in subsequent papers. He described expert systems which are used for information retrieval, but pointed out that most of these applications have not reached an extended commercial use, and are still in research and developmental states.

The next topic discussed was the use of CD-ROM in information retrieval, including the advantages and shortcomings of their use. Information stored on CD-ROM has a growing importance in the database market, and the number of databases on CD-ROM will increase dramatically. The use of CD-ROM is increasing in the use of parts catalogs for equipment and machines, as well as for maintenance manuals. Another increasing use of CD-ROM is as storage media for multimedia databases, for storing text, images, graphics and sound.

Hypertext systems allow the users of information retrieval systems to identify the relationships among the logic records of a database, and to display the related information on a microcomputer monitor.

Non-boolean search strategies were discussed, including document vectors criterion, cluster analysis, method of fuzzy sets, probabilistic retrieval, and search by means of the "nearest neighbor". these techniques are still in developmental stages, but the number of applications is clearly growing.

Lastly, the author discussed Local Area Networks (LAN) which are communication facilities which link devices in a small area. LANs optimize the concept of online communications by sharing expensive hardware, sharing software and sharing communication modems.

PAPER 1 - LOCAL AREA NETWORKS (LAN)

This paper discussed the man-machine interface problems encountered during installation, implementation and post-implementation of the LAN installed at the United Kingdom's Defence Research Information Centre (DRIC) in Glasgow. The LAN was considered necessary to increase the computing capability at DRIC, and to introduce enhancement and automation of maintaining card index systems, maintenance of records relating to distributions and requests, and the manual preparation of despatch notes, receipts, address labels and information relating to distribution and special restrictions. During this process, technology was not a problem, but rather the human aspect, the man-machine interface. The majority of the problems encountered were easily rectified by software amendments. Some of the man-machine/human aspect problems encountered were that terminals were wrongly situated, unidentified tasks and functions were discovered, and user complaints about poor response times.

PAPER 2 GATEWAYS "INTELLIGENTS"

This paper discussed the advantages of a gateway; namely, only one contract necessary to access several hosts; only one automatic procedure to the gateway; no confusion between several similar languages; possibility of multibase and multihost queries; adapted dialogue (user profile); choice of the more relevant bases for a query; user friendly interface; and efficiency.

The interrogation of a database in natural language is considered very interesting; the paper discussed natural language query and linguistic processing, and monolingual reformulation. The author states that multilingual access is necessary because in many cases, a user needs documents that may be in two or more languages. Multilingual access is useful to provide access to a database in a language the user does not know (in this case the system makes a computer-aided decision for a manual or automatic translation), in a language the user can read even if he is unable to make a query in it efficiently, or in which documents in different languages are mixed.

PAPER 3 - USE OF EXPERT SYSTEMS AS USER INTERFACE IN INFORMATION RETRIEVAL

The author provided a discourse on the difference between user-friendly interfaces, which essentially provide responses to the queries of the retriever, and intelligent interfaces that translate the retrieval query into in-depth comprehension.

Intelligent interfaces are based on a refined representation of domain knowledge (the database is, or is completed by, a knowledge base). In order to permit a better return of hits in response to a query, a more in-depth analysis of the query and documents is necessary.

Multi-expert systems incorporate more and more complex hybrid systems, which essentially means the integration of single interfaces into a multi-architectured system. The expert systems are replaced by multi-expert systems formed by specialist modules and strategy modules.

The interfaces which permit the use of natural language to interrogate a database are considered the first stage in the steps to multi-expert systems. The second stage is the development of intelligent interfaces formed by a data base joined to a knowledge base.

The application of the techniques of elaboration of knowledge bases from the contents of the documents and the preparation of paraphrases by means of lexic or linguistic transformations are necessary to prepare for multi-expert systems.

PAPER 4 SEARCH STRATEGIES IN NATURAL LANGUAGE

The author discussed online searching problems, such as: what relevant databases exist; how do you access them (autodial, gateway); how do you retrieve information from them (terminology, search strategy); and what can you do with the retrieved information (post-processing)?

The author then prescribed and described methods for making online searching easy for end-users, with emphasis given to parsing and natural language interfaces to the databases (structural Query Language, INTELLECT, SAPHIR, GURU, SCISOR). Natural language interfaces to multidisciplinary bibliographic databases include CITE, OKAPI, ALEXIS/DIANEGUIDE, PLEXUS/TOME/MITI, and DGIS/STINET.

The author continued with a discussion of vocabulary control and thesaurus aids, emphasizing the bilingual NATO Thesaurus. The author concluded with an explanation of the Netherlands search strategy - called 'Intelligent Information Retrieval' - which appears to provide very good recall.

PAPER 5 NON BOOLEAN SEARCH METHODS IN INFORMATION RETRIEVAL

All search strategies are based on a comparison between the query and stored documents. At times, this comparison is indirect (when the query is compared with clusters) or direct (when the query is compared with documents within a context of a given document). Oftentimes the comparison is iterative in that the user provides feedback after a first comparison which will affect the next comparison.

Search strategies consist of Boolean search, matching functions, and serial search. The basic instrument for trying to separate the relevant from the non-relevant documents is a matching function. Cluster base retrieval is based on the hypothesis that closely associated documents tend to be relevant to the same requests.

Feedback and evaluation are necessary to improve performance of a system by taking account of past performance. Basic evaluation measures of search and retrieval are efficiency versus effectiveness, where efficiency is measured using speed and storage overhead, and effectiveness is measured using relevance.

PAPER 6 - HYPERTEXT AND HYPERMEDIA SYSTEMS IN INFORMATION RETRIEVAL

Current, up-to-date definitions were cited, including:

document - recorded information structured for human consumption;

hypertext - a system of computer-supported, non-sequential information processing;

hypermedia - multimedia dynamic links among units of information;

multimedia - multiple forms of information; and

hyperprogramming - the process of creating hypertext or hypermedia applications.

Traditional information access methods are fundamentally linear, that is, a unit of information is read or viewed from beginning to end, with the document designed to be accessed with a clear path through the information from beginning to end. On the other hand, hypertext systems may provide the user with an initial linear access method, but at any given location in the information, the user has the option of selecting one to many further references. With such hypertext systems, the end user can pursue data references by following a self-selected trail or combination of trails through the data.

Hypermedia and related technology can improve both formal and informal information transfer despite barriers of distance and time (asynchronous annotation of information nodes) or language (computer-aided systems such as SYSTRAN).

The main added-value of hypermedia systems to the STI community lies in the ability of hypermedia to handle the full spectrum of STI's pragmatic content, from data manipulation to video display.

Several case studies on hypermedia development were presented, including Experiment Documentation Information System (EDIS), Life Sciences Interactive Information Recall (LSIIR), Decision Support System Shell, Knowledge Base Browser (KBB), the Space Station Freedom User Interface Language, PROJECT EMPEROR-1, and Clinical Practice Library of Medicine (CPLM).

PAPER 7 - AUTOMATED INPUT INTO DATABASES: OCR AND DESCRIPTIVE CATALOGING

Currently, paper still remains the most important medium of information exchange. The input process of this paper into a bibliographic database is a critical operation: it comprises more than 70% of all the costs of document installation.

Hardware and software for automated input into databases is available; the transformation process from written material into a machine readable database format consists of scanning at the graphical level, optical character recognition at the character level, descriptive cataloging at the document structure level, and subject indexing at the content level.

AUTOCAT was developed to produce records for a bibliographic database; its prototype application environment is INIS (the International Nuclear Information System of the International Atomic Energy Agency, based in Vienna). AUTOCAT recognized information elements in the machine readable journals, and normalizes the information elements and enters them into the target records - both steps as stipulated by INIS rules.

PAPER 8 - DATA COMPRESSION TECHNIQUES

Data compression is becoming accepted as a means to reduce the volume of documents, to make better use of available resources like communication channels and disk storage.

Data compression is essentially a matter of modelling the source of the data, and is sometimes referred to as "source coding".

Data compression algorithms are divided into reversible algorithms which only change the representation of the data into a more efficient one, and non-reversible algorithms which make only an approximate representation of the original data.

Currently, data compression techniques for text compression are readily available on most workstations and personal computers. A large share of the algorithms are in the public domain and can be freely used.

PAPER 9 - COMPUTERIZED PROPERTY DATA FOR ENGINEERING MATERIALS - AN OVERVIEW

The Material Properties Database (MPD) is a collection of data items whose values correspond to various large scale properties, parameters or attributes of materials, and are critically evaluated or validated by experts prior to their being included in the database.

The genesis and development of the MPD required a vast effort, because the data is difficult to deal with. For example, engineering properties such as creep strength of aluminum, are not intrinsic properties; rather, they will change as the material is loaded and as it ages, etc. Also, a particular material may be known by various nomenclatures in various countries or communities. Hence, entries for these properties must include data about the data, that is, a set of data descriptors and other associated information that characterizes the individual data values.

Because of the cost to obtain and disseminate materials information, and because it is so vital to the manufacturing industry, materials information must be regarded as an international commodity. Hence, standards are being and have been developed to relate to the quality and reliability of data, database system management, system capabilities and data security and integrity.

PAPER 10 - FACILITATING THE TRANSFER OF SCIENTIFIC AND TECHNICAL INFORMATION WITH SCIENTIFIC AND TECHNICAL NUMERIC DATABASES

Numeric databases are collections of information and data, and contain both data and metadata, or textual information relating to the data. Scientific, technical and engineering databases comprise the second highest subject category of all numeric databases next to business databases.

In an effort to better serve the scientist and engineer, the U.S. Department of Defense Defense Technical Information Center (DTIC), through its Defense Gateway Information System (DGIS), provides the end user with an access mechanism to databases, and through its Multi-Type Information and Data Analysis System (MIDAS) will provide a capability for the end user to process bibliographic information and numeric data.

DTIC conducted an S&T Numeric Database Technology Assessment to more thoroughly understand the information and data resource needs of the scientist and engineer, as well as the computing environment in which they function and operate. DTIC has also identified scientific and technical numeric databases throughout government and industry. The information gleaned during this assessment will determine the extent of the investment DTIC will make in providing expanded services to their users.

SECTION 2

COMMENTS, CONCLUSIONS AND RECOMMENDATIONS

Rather than dwell on an evaluation of each presentation, I will just make some generalizations. The Keynote Address served as a condensation of the status of the current, state-of-the-art techniques and technologies that were to be presented at the meeting. It was an excellent introduction to the meeting.

The subsequent papers presented reports of experiences, curren: assessments of developing technologies, and assessments and evaluations of current systems, software and hardware. Each topic was well understood by its author. Time constraints at times prevented an in-depth total coverage. However, the question and answer periods accommodated pressing questions and problems. The speakers were well prepared and obviously experienced in their areas of specialisation. The technology discussed was relevant to the interests and objectives of the TIP Specialists Meeting. However, the papers were not totally in concert with the stated theme. for example, the following omissions were noted:

"...absence of suitable standards..."
not often mentioned except in passing.

"...short life associated with..."
not discussed at all.

"...successes and failures..."
successes were highlighted; few failures were mentioned.

"...finding practical solutions..."
not many papers did.

The past several TIP meetings have dwelt on technology, existent and emerging, that will have an impact on how information resources will be allocated and managed. This is all well and good, but is this new technology which is being developed, developed with the user, both the intermediary and end user, in mind?

It is well to be aware of current developments. However, I feel that we must every so often come back to basics and look at what we have done, where we are, and what will be our future; keeping in mind that all developments must be in concert with our users and their requirements.

The TIP Terms of Reference state that the Panel is concerned with all aspects of the management of scientific and technical information as an integral part of the aerospace (and defense) research and development process. Timely, accurate and relevant STI is critical to the R&D process; it is an incredibly valuable resource that directly affects the cost of performing a technical task, the quality of the results, and productivity.

Unfortunately, during the past few years, STI program managers have been battling budget cuts, coping with personnel cuts and losses, acquiring new equipment, etc. The relationships between R&D managers and STI managers have loosened; more and more they work as separate communities, with the STI community serving a passive role by responding to service requests of the R&D community.

STI Programs must refocus and concentrate on how better to support the R&D community, as well as how to support scientific and technical productivity. This apparent gap between R&D managers and the STI managers must be filled; information specialists must be actively involved in all stages of R&D. This participation must not be a passive "Don't call us we'll call you", but the result of active membership on the R&D team.

There are several trends emerging which have a significant impact on the conduct of science, research and development and the corollary management activities, and which dictate that generic issues of STI be addressed. The trends include the use of information technology, the growth of interdisciplinary research, and an increase in international collaboration.

Information technology, which has dramatically changed the conduct of research, has brought forth a need to better understand and manage its exploitation. Computerized instruments gather data many orders of magnitude greater than previous methods. Telecommunication capabilities link researchers to computing facilities with vast capabilities and with data sources not constrained by geographical location. Data are available, not only in computerized databases, but also from sensing and other data gathering instruments. New analytical approaches are possible through graphics, color enhancement, animation, and other visualization techniques. With this ever growing capability, there is a need to help teach researchers to better use it, to develop better ways to store, retrieve data and to maintain its integrity, and to determine how to assure intellectual property rights in an electronic network.

Many of the significant research challenges today are interdisciplinary in nature, which requires expanding the circle of collaborators, as well as the range of information sources. A network of communications links will soon develop worldwide, to link personal computers, work stations, data bases, peripherals, and information utilities. Information systems will become transparent, and will facilitate the flow of information and meaning among people. Consequently, we will be able to focus on content not technology. Responsive expert advice, information, and solutions will be at our fingertips; we will find ourselves receiving more stimulation and excitement from the systems than the energy we put into them. We will become more purposeful, growing, and professional than we are now.

Notwithstanding these communications networks and large databases, the different methodologies, vocabularies, and cultures of individual disciplines create obstacles to efficient information exchange. Systems need to be designed to accommodate users who were not immediately involved in the original research. Merging existing data collections from different fields to perform analyses creates new problems. It becomes extremely difficult to compare data that were derived using different techniques or approaches. Contributing to this problem is the lack of standards for data exchange formats which hamper the building of these multidisciplinary databases. The bottom line is that we must be prepared to import external information to support the internal R&D process, assure real-time delivery of information to support the transfer and transition of technology within the R&D community, and be able to export some results to remain competitive in the R&D arena, as well as to provide visibility to the organization.

These problems are further compounded by the growing internationalization of science. STI is being produced, enhanced, and stored around the globe. Single countries in some cases are acknowledged leaders in select scientific and technical disciplines. Many of the major research efforts involve worldwide data collection. Not only are a variety of disciplines involved, but scientists from around the world are participating in these efforts. The users in these projects are distant geographically as well. Global economies dictate that every effort be made to reduce unnecessary product and service development cost. Communications networks facilitate the exchange of ideas and access to remote databases, but there is still much progress that needs to be made in making systems more transparent and in developing common protocols.

Hence, the pace of data collection, the growth of international approaches to research, and the tendency to cross traditional disciplinary boundaries all cast a new perspective on earlier STI issues, and raise new challenges for effectively providing critical information to the end user.

The issues to be addressed and resolved are numerous, including the transparency of access to vastly expanded and distributed electronic resources; merging data from numerous sources; greater data validation; closer cooperation between the user community of scientists, engineers, and managers and the information system designers; the long-term viability of electronic data; and expanded resource commitments to support technologically advanced information systems; archiving large scientific databases; what STI should be retained, where datasets should reside; what formats should be used, how can they be physically maintained; and how to reduce dependence on specific hardware and software.

Notwithstanding the above issues, it will also be necessary to better understand the knowledge transfer process. It will be necessary to establish a research agenda to address these and other issues related to STI. Not only must information managers, but policy makers involved in the science and technology programs as well, need to understand the relationship of STI to the R&D process, namely, that knowledge transfer is an inseparable part of R&D. Innovation is a complex process composed of multiple and interrelated systems. A better understanding of knowledge diffusion by policy makers, R&D managers, scientists, engineers, and information specialists should result in better defining policy and programs that will enhance the productivity of the R&D community, and in turn enhance competitiveness.

As STI concerns move beyond the parochial interest of particular disciplines, as linkages occur with the networking community, and as the trends toward interdisciplinary research on a global scale become more pervasive, an expanded R&D user community is developing. The user community must voice legitimate concerns about both technical and policy issues associated with STI. The user community must identify common concerns about STI access and in building systems that will accommodate the needs of future government scientific and technical initiative.

We are in the dynamics of technological pull versus administrative lag. Administrative lag retards the development and use of the new information systems and technologies. The industrial age from which we are departing needed us to be interchangeable cogs in a machine, turned-off and emotionless, mechanical, routine, controllable, and consistent. The new information age into which we are entering needs us to be growing, experimental, creative, enthusiastic, risking, and taking initiative.

CONCLUSIONS/RECOMMENDATIONS

* STI must be considered as an R&D resource, essential to the continued success and innovation of the R&D community. Not to be overlooked is the fact that STI has costs: costs in collection, internal and external communications, processing and storage, archiving and disposal, and in skilled staff used in all of the activities above. It is also noteworthy to mention that although STI is used mainly by the scientists and engineers of the R&D community, it does have value and is required at the policy level, as well as at the managerial level.

* STI management means more than simply developing more sophisticated information transfers systems; rather, it means providing the means to exploit both internal (corporate) and pertinent external (other governmental/ industrial/ foreign) information to meet the requirements of the R&D community.

* Practical steps must be taken to improve the quality, timeliness, and accuracy of information which will have an impact on the efforts of the R&D community. By recognizing problems and taking appropriate action to correct them, information handling costs can be reduced. Given the size of expenditure on information handling, even small improvements in the efficient use of information can result in very large potential savings.

* Effective use of information adds value to all the activities of the R&D community. It means improved quality of information for more effective planning; more effective and efficient discharge of functions and higher quality of service; more accurate, more cost-effective information; reduced expenditure on the collection, communication and storage of unnecessary data; and a better focused information system investment.

* However, it is only in close concert with the R&D community that we can make the most effective use of information. It is only in concert with the R&D community that we can identify and specify the needs for information (including its content, quality, and timeliness); identify the most appropriate sources of information to meet these needs; identify the most appropriate mechanism for the delivery of this information; and establish procedures to allow data from many sources to be brought together to provide information at the point of need. In short, it is only with the help and cooperation of the R&D community that STI Programs can provide information services which are easily accessible, and allow users to find the information they need with the minimum difficulty and minimum intervention by skilled specialists.

* The starting point for information management must be an understanding of the users' business, its aims and objectives, and how these are translated into the functions it performs. It is then possible to derive or work out the total information needed to carry out this mission. It is important to note that the product which results from processing the required information is very important, and should employ language familiar to the users of the information.

* Through dialogue and support of the R&D community the differences between information need and provision can be investigated, and this investigation will determine where it is necessary to make up the deficiencies or dispose of the surpluses. The choice of delivery systems depends largely on who needs the information, how quickly, how frequently, and what they do with it. Exploitation of the information stock also depends on knowing what is available, and on being able to identify whether it offers a contribution to the requirements. The tasks are all continuous, requiring constant or periodic review, which is best done during R&D planning stages.

* STI management must become a part of the accepted culture of the R&D community, but it cannot become so unless adopted and accepted by it. A start should be made now to integrate one's STI Program into the R&D infrastructure, including funding and operational control. Within the R&D infrastructure, we must obtain management commitment, review and produce policy reflecting our organizational status, allocate responsibilities and set to work on implementing the true requirements of the R&D community.

KEYNOTE ADDRESS

A. del Rey
I.C.Y.T. (C.S.I.C.)
c/Joaquin Costa 22
28002-Madrid
España

In this keynote adress we intend to give a brief overview about almost all the topics which will be discussed during this Meeting.We will begin speaking about the use of expert systems in the information retrieval field. Since many years ago expert systerms are being used for information retrieval, but until now most of these applications have not reached an extended commercial area and are limited in most cases to the use in a University or investigation Center.

The expert systems can be used in information retrieval with the following purposes:
-Selection of a suitable Database to carry out a search, among a set of Database available from a set of Hosts.
-Formulation of a information request in natural or controlled language.
-Translation of the information request to a search strategy and automatized searching.
-Displaying of some search results and obtaining from them new search terms to be included in a new search strategy and iteration of the search process with this new search strategy to reach the recall and precision requested by the user.

Which are the characteristics of the expert system to fullfill these requisites? These characteristics are the following:
a)ability to carry out the parsing of the user information requests.
b)ability to control the man-machine interaction.
c)program ability to explain its capabilities, unfitnesses, what it is doing and what it does for.
d)ability to identifiy an object from a description.
e)ability for heuristic learning(by trial and error).
f)ability to ascertain the user knowledges on the search topic.
g)ability to correct user errors.
h)to operate in user friendly mode.
i)use of tutorials to guide the user.
j)suitable system response time and
k)to be able to help user in the search strategy preparation.
We can add to these requisites these other ones:
1)ability to understand information requests in natural language
2)ability to use the automatic weighting of search terms and to calculate the global weight of a reference
3)iterative search of references like previously obtained records.

To carry out all these requisites the expert system must operate in the following way:the expert system asks the user for the concepts dealing with his information need and then asks him for the search terms. Then explains to the user the various operators used in the search process and show the user the initial search strategy.

Later the expert system asks the user for the wanted recall and precision and to classify the search in a broad subject classification to select the Database set to be used in the search.Then the search is carried out and some records are displayed to the user to decide the iteration of search process, to broaden or to restrict the number of records obtained as results or to take out new search terms of the displayed records. After this the system proceeds if neccesary, to iterate the search process until a good search result is reached.

In short we can say that expert systems in information retrieval have the following objectives:
.to give information over-
to help the user for-
to carry out instead of the user-
to avoid the user the knowledge of one or some of the following processes:
a)preliminary processes(telecommunications and host selection,telecommunications use, host access or DB selection)
b)search process(search strategy planning, search terms selection, iterative search)
c)handling of records recovered in the search
d)ancillary processes(errors correction, search data recording, etc.)
e) translation of information requests in natural language into a search strategy.

The second topic to be discussed in this keynote will be the use of CD-ROM in information retrieval.
The high storage capacity of CD-ROM, between 550 and 600 Mbytes, made very interesting its use in automated information retrieval, since a bibliographic Database with 1,000,000 records can be stored in a little number of CD-ROMs (between two and six in accordance with the records average size).

The use of CD-ROM Databases have the following advantages:
-to avoid the use of telecommunications access to a main computer to do an online search on large Databases, in this way the PTT and host payments are not neccessary
-possibility of extending in time the search without any of the above mentioned costs
-to use the CD-ROM Database only is neccessary the payment of an annual subscription, receiving a large fragment of a Database or all the Database.

Among the shortcomings are the following:
-the access time is very high in comparison with the online one and this gives rise to a slower information retrieval process. If we take in account the need of consulting several CD-ROMs to do a search, it will be possible to use 40, 50 or 60 minutes in doing a search in CD-ROMs, if the number of file years to be searched is enough high.
-the time delay between the publication of a document record and its appearing in CD-ROM is somewhat high in most cases(around 1 1/2 - 2 months after the appearing of the record in an online Database).This delay must be acceptable for a great number of users, but for some is unacceptable. Anyway this delay is becomig shorter than before and currently many CD-ROMs have information available only one or two weeks after than online.
-another shortcoming is the high price of most Database in CD-ROM which only makes profitable the use of this type Database storage when a certain number of searchs are done yearly. And therefore if we need to acquire many Databases on CD-ROM its costs can be very high.

Now we will give some data to discusse the increasing expansion of the use od CD-ROMs in information retrieval:
-in a study carried out in 1987 it was noted that in 1986 were put in the market 19000 CD-ROM units, in 1987 would be put 50000, in 1988 about 137000 and in 1989 c.a. 597000, that means to multiply by 30 the sales in four years.
-in the same study the numbers of Databases on CD-ROM were the following: in 1985 5 DB, in 1986 25 DB, in 1987 125 DB and in 1988 c.a. 210 DB.
-in "The CD-ROM directory", edited by TFPL there were 350 DB in 1989, 715 DB in 1990 and for 1991 the figure is c.a. 1450 DB,that means a multiplying factor of 500% for the last three years.

From all these data we can conclude that the the CD-ROM stored information has a growing importance in the Database market and that in the following years the number of DB on CD-ROM will increase strongly. On the other hand taking into account the decreasing prices of CD-ROM subscriptions and the increasing number of Database on CD-ROM, we can suppose that in in the next 8 or 10 years, the use of online Databases will be greatly reduced, being used these Databases only when the CD-ROM counterpart is more expensive than the online one or, of course, when the Database is not on CD-ROM.

Other increasing application of CD-ROM is it use in the elaboration of parts Catalogs for different equipment and machines or for equipment maintenance manuals (e.g.for commercial planes, automobiles, etc). In these cases the large capacity of CD-ROM is used to store graphics and images in digitalized mode.

Another application of CD-ROM is as storage medium for multimedia Databases, for storing text, images, graphics and sound, this application is growing increasingly.

Another very interesting feature of CD-ROM in the field of information storage is the use of this disk in erasable form, i.e.erasable CD-ROM. At the present time the storage medium of this type which have reached more commercial diffusion is the magneto-optic disk; the writing process is the following: the binary 0 and 1 are stored in magnetic form by means of a high power laser beam which heats the base material at a temperature at which the orientation of magnetic particles is easily changeable by means of a weak magnetic field, afterwards the laser beam is released and the new magnetic particles orientation is "frozen", ending at this moment the writing phase. In the reading process it is used a low power laser beam and this, when incides on the disk surface is polarized in one direction if there is a 0 or in the opposite, if there is a 1.

The erasing process is carried out by heating the base material with the high power laser beam and applying on the corresponding magnetic particle a magnetic field opposite to the initially applied.

These magneto-optic disks and other with different physical basis are used increasingly to store large data amounts in a little volume and we can suppose that they will replace the magnetic discs in a near future, for instance to store personal Databases of large size.

Currently is increasing the use of CD-ROM networks which allows the simultaneous access of many users to a CD-ROM Database. The use of these optic networks is linked with the use of local area networks and in this way allows the access to data from different access points. In this way we can say that the "online" access to CD-ROM Databases is beginning, with the consequences that this fact means.

The following topic to be discussed will be the use of hypertext in DBMS(Data Base Management Systems); this development appears to be very interesting in a near future.

Since 1987 the possibility of using hypertext systems has been considered, to allow the users of information retrieval systems to identify the relationships among the logic records of a Database and to display the related information in a microcomputer monitor. We can mention for instance the IR system Rivage, which has an operating scheme based on hypermedia and allows the interactive browsing of a images Database stored in video-disc.

Hypertext technology can be used for the following objectives:
-storage and management of non-lineal documents
-storage and management of a document Database in hypertext
-management of the semantic structure of concepts (paradata) together with the management of a links network, which can serve for a thesaurus management and the management of the documents containing those concepts.

In this frame each hypertext node must contain a full document or a part of a document and the network of links among nodes is used to connect structurally the different parts of documents and to connect documents with semantic similarity.

To get the management of Databases by the hypertext system is neccessary to define the following types of nodes and links:
-text nodes (capables of containing text fragments)
-topic nodes(capables of containing the semantic description of a concept)
-structural links (allow the non linear documents structuration)
-semantic links (allow the connection of topic nodes)
-connecting links (manage the relationships between the concepts appearing in thesaurus and the documents of the hypertext Database)

In brief: a Database managed by an information retrieval system consists of two components:
-the set of documents or their surrogates and
-the indexing terms (paradata)
and in a hypertext system, there are two components:
-the set of documents or their surrogates and
-a links network connecting the documents with a semantic or stuctural relationship which is equivalent to paradata.

Partial matching criteria must be used in the search of records in a hypertext Database, between the documents search models and search strategies.

Each topic node (containing a thesaurus term) is related with the set of documents pertinent for this term and the links network among topic nodes support all the relationships among the terms used in the thesaurus.

The following search types can be done:search of character strings; non sequential browsing of document records following the links between concepts belonging to the documents, and search from a pertinent record, previously found, following the weave of hypertext links.

The links between nodes can be as follows:
-links between descriptors and indexes, which can link descriptors with a thesaurus, a hyerarchic index or a permuted term list
-full text links, to link similar documents and
-citation links, to link two or more documents cited in another one

Currently is being develoveped a series of hypertext applications to the management of hypermedia databases (with records formed by a combination of audio, video and text).

An example of online application of hypertext to Databases management is the Hyperline of ESA-IRS. This facility is an information browser that allows concepts and reference browsing and carries out the semantic association between user searched concepts and concepts stored in the information retrieval system.

Hyperline allows to integrate two basic elements of information retrieval: the document browsing and the navigation through concepts and adopts the computer interaction with browsing and concept association.

In which way is Hyperline elaborated? Documents in bibliographic Databases are indexed and transformed into records, being also classified. At the same time as classification, a knowledge base from the involved concepts and their mutual relationships is built. This knowledge base is introduced in the computer with the bibliographic records; the knowledge base is used and explored by Hyperline, allowing the concept navigation and the reading or browsing of references in any navigation moment.

It has been developed a new information retrieval model with a two levels architecture; the set of relevant records appears in the first level and the semantic related concepts are put in the second one. The first level is managed by the I.R. system, and the second one has been designed as a conceptual interface between the user and the records set.

In the interface man-computer the following functions must be included:
semantic association; concept navigation forward and backward; sequential and associative reading of references; history of the interaction and support for query formulation.

We discuss each of these functions in the following:
-semantic association: the purpose of semantic association function is to give the user an entry point into the concept network stored in the system. The user write his query concept in natural language words and that is put by the system in a list of semantic rélated concepts, which is a part of the knowledge base which manages the system. In this way the user receives a system answer regardless the terms used in the query. The list of semantic related concepts enables him to initiate the concepts navigation.

-Navigation: this function present to users the possibility of browsing the semantic concepts structure which represents the information contents of the bibliographic records.

-References reading: in any time, during the navigation by data structure, the user can read the bibliographic references containing the term or concept which is being examined in the semantic network.

-History: The history function keep details on the history of the user-system interaction during an Hyperline session; it displays all functions executed during the navigation process.

-Support for query formulation: At any time during concept browsing, concepts can be selected and put aside for subsequent use in Boolean query formulation. This allows an intimate interconnection between the classical Boolean searching and hypertext browsing.

Another topic to be discussed will be the Nonboolean matching criteria.

One of the more serious shortcomings that the use of Boolean logic operators has as match criterion is the impossibility of classifying the records obtained as search result according to its relevance for the user; in other words: it is impossible to rank the records putting first the records in which the search terms are very important and secondly those in which the search terms are not important. This is due to the unability to assign a grading of terms importance in the document (or record) retrieved, since the assignement of indexing terms to a document is completely binary (if the term is important it is assigned and if it is not important it is not assigned).

A lot of other matching criteria have been developed to avoid this shortcoming and the more important are the following:
-document vectors criterion, in this method the indexed documents are represented by a set of document vectors; each vector is a set of concept numbers (codes) with weights, the concept number represents the indexing terms assigned to each document and the weights are the relative importance of each term in the document. The information request in natural language is translated in a similar mode to a search vector and the retrieval process is carried out by comparing the search vector with all or some of document vectors. Then document vectors are ranked by descending order of coincidence with the search vector. The search vector is modifyed by relevance feedback and the search is itered until obtaining good results.

-Cluster analysis, the document clusters are prepared by comparison between the indexing terms of a document and the indexing terms of the other ones and clustering those documents whose indexing terms are similar. For each cluster a representative element is chosen, named centroid vector and the search is carried out in two steps: in the first the search strategy is compared with all the centroid vectors and in the second one the search strategy is matched against the individual documents of clusters with centroids very similar to search strategy, found in the first step. This step search can be broadened to three or more steps by grouping centroid vectors in broader clusters of greater coverage. In this case the search begins by the broadest clusters and then is carried out with the more specific clusters.

-Method of fuzzy sets, in this method a fuzzy set of document identifiers is assigned to each index term, in this fuzzy set the grade to which each term belongs is given by a weight between 0.1 and 1, and when the term does not belong to the set its weight is 0. With the fuzzy sets all the Boolean logic operators can be used but it is necessary to lose some axioms of that logic to obtain consistent search results.

-Probabilistic retrieval, a system of probabilistic search can begin a search by assigning numeric values of probability or uncertainty to indexing terms and by using the probability rules can obtain the probability of pertinence of a document for a search topic. These pertinence probabilities, obtained from the document indexing terms govern the information retrieval decisions of the system. In these systems the Boolean relationships are lost.

-Search by means of the "nearest neighbor":
in this method a matching is carried out between the general set of search terms and the

indexing terms of each document and the selected documents are ranked by descending order of likeness with the request.The "nearest neighbor" is the document with a set of indexing terms most similar to search strategy that has been found in a certain moment of the search.When a document is found, which is more similar to search strategy than the actual "nearest neigbor", that document becomes the new "nearest neighbor".

-Among the probabilistic retrieval methods appear the weights method, in which weights are assigned to search in each document or/and in the search strategy.This method used in combination with the Boolean operators one, enables us to avoid some shortcomings of the Boolean method and has some implantation in Database market.

All of these methods have only a low implantation level in the daily rutine of information retrieval; this is due on the one part to the low impact of these investigations on the commercial services of information retrieval and on the other part to low knowledge level in this field of the Hosts staff members about the investigations in the field and its theoretical basis.

Among the advantages of non-Boolean matching criteria are the following:

-it is not necessary to prepare a Boolean search strategy, it is enough to present to the information retrieval system a set of search terms that we want it to appear in the documents to be retrieved.

-the weighting coefficients are easy aplicable.

-the feedback of search terms from retrievced records is easy.

-it is possible to use a search terms amount larger than in the Boolean search method.

-a flexible Boolean logic can be used.

-the relationships between search terms and documents can be expressed.

And on the other hand we can add the following to the shortcomings of the Boolean method:

-it is necessary a very good indexing

-it is necessary a correct use of the Boolean logic, which is difficult many times.

Finally we will speak about the use in the Library environment of Local Area Networks (LAN).

We can define a LAN in the following terms: it is a private owned communications facility which link devices in a small area.

We will comment this definition:

-a LAN must be fully owned and operated privately by the Institution which funds its activities.

-it is a communicatons facility that allows devices to exchange information; several of the components of a LAN must have idependent intelligence and processing power.LANs can transmit: data, video, images, audio messages and facsimile.

-a LAN link devices, a device linked in a LAN must be able to communicate at least to another device on the network. The linked devices can be: CPU, command and control systems, dumb or intelligent terminals, fax systems, interactive video, peripheral devices (tapes, disks, etc.), telephones and message transceivers.

-it is a facility for a small area (with a maximal separation among devices of about 2 km)

The LAN components are the following: the cabling system, workstations, servers, interface units and network software.

There are three systems for cabling:twisted pair wire, coaxial cable and optic fiber cable. The twisted pair wire has a main weakness:it is susceptible to noise, however this drawback can be minimised with a proper shielding of the cable. Its transmission rate is the lowest, only reach from 250 Kb/s to 2 Mb/s in the baseband.

Coaxial cable can achieve higher transmission rates, about some Mb/s without signal regeneration, it also allows greater distances than twisted pair wire and a greater number of attached devices.

The highest performance is attained with the optic fiber cabling, that allows transmission rates of some Gb/s, has low weight and is more noise resistant than the other two cabling systems. By this cable can be trasmitted voice, images, video and data.

The workstations are microcomputers used to access or manipulate data.

The servers are microcomputers which provide access to shared resources; for each shared device or set of devices an associated server must be contacted before use. There are systems with independent servers and other ones in which the server function is allocated to some of the workstations.

The interface units, which allow the logical connection between the computation devices.

The LAN software: a LAN needs software to run and perform all its functions. There are three types of software: the system software, which manages the hardware and allows other software to operate using the host hardware; the network software which allows the interconnection between applications and the network and the applications software (wordprocessing, DBMS, etc.).

Topologies: The LANs have three basic topologies: star, ring and bus/tree. A star is a group of connected devices (nodes) served by a central device. A bus/tree is a multiple access broadcast medium, the bus being a special case of tree with only one trunk. The ring is a closed bus with each node attached to a repeating element.

The most common accessing protocols for LANs are the following:carrier-sense multiple access with collision (of messages) detection (CSMA/CD), carrier-sense multiple access with collision avoidance (CSMA/CA) and token passing. In the CSMA/CD protocol a network node transmits a message after listening to the network to make sure it is not busy and then begins to transmit. In the CSMA/CA protocol message collisions get detected by the sending and receiving nodes. Token passing is a deterministic accessing scheme, since each node is guaranteed access within a set period of time and the access is controlled by a token circulated among all nodes at a constant speed.

LANs optimize the concept of online in several ways: make easy online sharing of expensive hardware; the software can be shared online without physical transporting diskettes form one micro to another and the communications modems can also be shared.

The main reasons for networking microcomputers in Libraries are the following:
-to access common data (e.g. OPAC, acquisitions and serials information)
-to share expensive devices (hard disks, high quality printers, plotters)
-to share software
-to allow electronic mail among departaments and patrons and
-uploading and downloading from other systems (data banks).

The LANs are also used for the following reasons:
To improve communications and to share equipment and software.
LANs can manage relatively large data files and support a large number of simultaneous users on a locally controlled system and also provide for the efficient use of resources trough shared peripherals, such as printers, plotters and software
LANs provide for system data security trough centralized backup and for data integrity trough shared use of a central file, it also may contribute to improved communications.

Finally we will consider some of the issues to be anlyzed when considering the installation of a LAN:
-are there sufficient requirements for using a LAN in your environment?
-Is LAN hardware and equipment available to you from the financial and physical viewpoints?
-will the LAN be reliable?
-is there growth potential for the LAN?
-is security and control on a LAN an important issue?
-how will traffic volume on the LAN affect your access?
-what kind of speed on the LAN will provide the required response time?

I hope to have given a suitable global view of some of the more interesting topics presented at this Meeting with the above considerations.

LOCAL AREA NETWORKS (LAN)

G. FARQUHAR
Defence Research Information Centre
Kentigern House
65 Brown Street
Glasgow
Scotland
United Kingdom

1. SUMMARY

The discussion is confined to the Man Machine Interface problems encountered during implementation and post implementation. Various development aspects are considered, commencing with the definition of the users' requirement, as distinct from the users' wishes, to the provision of adequate post implementation support.

The LAN installed at the Defence Research Information Centre(DRIC) during 1987 and subsequently enhanced in 1988 is taken as the model for the discussion.

The aspects considered are:

DRIC's working procedures - immediately prior to installation of the LAN in April 1987.

Phased implementation - of the LAN and associated consultation procedures.

Phase 1 - Provision of Information Retrieval facilities, for scientific staff.

Phase 2 - Provision of a Document Movement Control System.

Post implementation review - review of requirements, problems identified, user reactions.

Current Activities - endeavours made to achieve a satisfactory system performance level, including changes to, computer processing pattern, working procedures, supply of office furniture.

For each of the aspects listed above details of problems encountered and solutions implemented are given.

2. INTRODUCTION

2.1 General

DRIC is the MOD's central deposit and dissemination point for defence scientific and technical literature to the UK and overseas Defence Community. In March 1986 it moved from St Mary Cray to central Glasgow, where it is now located along with a number of other MOD branches in a modern office block. It was formed in October 1971 by the merger of the defence component of the Technology Reports Centre (TRC) with the Naval Scientific and Technical Information Centre (NSTIC). DRIC is part of the Assistant Chief Scientific Adviser (Research) (ACSA(R)) Organisation. Mr M R C Wilkinson is Head of DRIC and he reports to Director Research (Technology). In order to discharge its remit DRIC is organised as follows:

HEAD DRIC
MR M R C WILKINSON

HEAD GROUP 1 DR E H MORGAN	HEAD GROUP 2 MR G FARQUHAR
Document Abstracting & Literature Searching	Information Technology Policy & Strategy
Document - Recording - Requests - Acquisitions	Provision, Development and Maintenance of IT facilities
Document - Distribution - Downgrading	User help desk
DRIC Publications	Administration
SDI Service	Document Registration
Document Translations	Stockroom, including document despatch

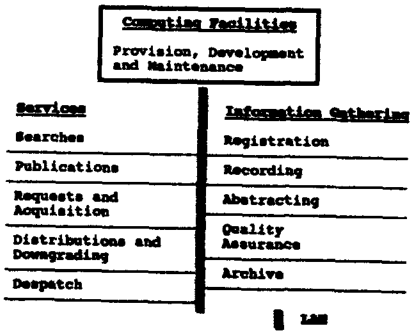
DRIC's total complement is 57.

2.2 Computing facilities

The current configuration, is based on a Digital Equipment Company (DEC) VAX 6310 computer and consists of:

32 Mb of memory
3.2 Gb of magnetic disk
3 Laser printers (desktop)
3 Dot matrix printers
1 Magnetic tape unit

The functions currently supported by the LAN are depicted in the following diagram:



The computer databases hold over 250,000 records of documents available from DRIC and dating from about 1970. The DRIC holding however, consists of approximately 600,000 document titles dating back to around 1940. The databases are currently accessed via a LAN, supporting 40 video terminals, 2 bar code label printers and 18 bar code label readers. The LAN consists of a mix of Fibre Optic and Ethernet Cables.

3. DISCUSSION

3.1 DRIC - before April 1987

Prior to April 1987 DRIC's databases were held at an MOD Bureau facility in London. In order to update these databases, data was extracted from the relevant documents and entered on computer input forms. These forms were encoded into machine readable format by DRIC's own data preparation staff, on an in-house facility based on a GEC 4080 computer. The data was then transferred, via the postal services, to the bureau.

The information retrieval software in use at the bureau was limited in its capability and lacked flexibility.

The availability of computing facilities was restricted to scientific staff involved in, information retrieval services and publication of DRIC's announcement bulletins. Access was by a single terminal shared by ten scientists. All output generated by database searching was printed at the bureau and was delivered by post to the scientist concerned. At this point the scientist got out the scissors and glue pot (cut and paste) to produce an acceptable result for the requester. This naturally caused considerable delay to the supply of information.

Acquisitions, distributions and the supply of documents to requesters were handled manually. This entailed:

The maintenance of an enormous card index system containing document, receipt, movement and bibliographic data.

The maintenance of manual records relating to distributions and requests.

The manual preparation of, despatch notes, receipts, address labels and information relating to distribution and special restrictions.

At this stage there was no automated procedure for registering receipt and movement of documents. Documents arrived in DRIC and their existence was unknown to the system until they were entered in the bureau database many weeks after arrival. These conditions led to a number of minor security breaches causing the deployment of scarce resources to investigate.

3.2 Preliminary Study

In 1984/85 it became necessary to consider the replacement of DRIC's computing facilities.

This was due to:

The age of the in-house computer used for data capture.

Changing factors at the bureau.

The age and source language of the retrieval software in use at the bureau.

The impending move to Glasgow.

The first step was to establish the scope of the requirement before examining the various hardware and software options available. A small working group was therefore set up consisting of representatives from DRIC's, Information Technology (IT) Section - *the system providers*, and Publications and Technical Enquiries Sections - *the users*. The group also included a representative from Director General Information Technology Systems (DGITS). DGITS is the MOD directorate responsible for all aspects of IT Policy and Standards, including, Technical appraisal and Procurement.

A representative from the Central Computer and Telecommunications Agency (CCTA) was also involved with the working group. CCTA is part of HM Treasury.

The group was tasked with:

Establishing the requirement.

Sizing the requirement in terms of;

Input and Output data volumes.

Fixed data, ie database volumes.

Database searching.

Volume of printed output relating to Document, announcement, supply and distribution.

The number of Video Terminals needed by the various user sections.

Identifying commercially available software packages capable of meeting the requirement.

Identifying hardware capable of supporting the sized requirement and the software package judged to be to most suitable.

Producing an implementation plan which would cause the least disruption to DRIC's customer service.

3.3 System Selection

The working group established the system requirement as consisting of two separate functions:

Phase 1 - Provision of Information Retrieval and Document Announcement facilities.

Phase 2 - Provision of a Document Movement Control System.

This led to the decision that the Information Retrieval and Document Announcement facilities should be implemented first. It was anticipated that this would be reasonably straight forward since the relevant data and these facilities, in a limited form, were already available at the bureau.

The sizing exercise indicated that, based on the assessed volumes of Input and Output data and projected database size, there would be a need for 1.8Gb of disk storage for data and system files at the system installation stage. It was also estimated that this requirement would increase by another 2Gb within 3 years of installation.

The overall number of Video Terminals required was estimated at 37 (10 for Phase 1 and 27 for Phase 2).

The working group identified a number of commercially available information retrieval packages. An evaluation of six packages reduced the number to two which were considered most suitable for the requirement.

The package eventually chosen was Computer Aided Information Retrieval System (CAIRS) supplied by Leatherhead Food Research Association (LFRA).

It was decided to invite LFRA to tender for the supply of the software package and suitable hardware to support the requirement. This required DRIC to draw up a full Operational Requirement (OR) for agreement with DGITS and CCTA. This document is the basis on which a supplier is invited to tender and sets out all mandatory and desirable requirements of a project.

After much discussion and modification the OR was agreed and in October 1986 LFRA were invited to tender for the project.

The implementation plan produced by the working group required that the system be introduced in two phases:

Phase 1 - Provision of Information Retrieval facilities for scientific staff.

Phase 2 - Provision of a Document Movement Control System (DMCS).

Phase 1 would begin with the installation of the hardware, software and conversion of existing data held on the bureau computer. Phase 2 would be designed and produced in-house by DRIC's IT section.

The target date for the start of phase 1 was set for 1 May 1987 while the date for phase 2 was set for April 1988.

By the end of 1986 LFRA had submitted their proposal to supply the CAIRS software package, additional customised conversion and print software and hardware to support the system. The hardware proposed was a DEC VAX 8200 computer.

3.4 Phase 1

System Prototype

The databases at the bureau had been designed in the early 1970s for a computing environment supported by International Computers Limited (ICL) and had been modified a number of times since then. There was therefore a need to convert the data from ICL to DEC and then to CAIRS format. This being the case the opportunity was taken to;

- remove obsolete data elements.

- introduce additional data elements.

- rearrange the order of the data elements.

DRIC, with the assistance of a consultant from LFRA, implemented a prototype of the Phase 1 system on a PC based version of CAIRS. The prototype was most useful, enabling both IT and Scientific staff to trial and modify the system before full implementation.

Indeed it was at this stage that DRIC made a fundamental change to the system requirement,it was decided that Phase 2 - Document Movement Control System should have its own separate database.

The system requirement as originally specified, proposed:

> That the bureau database format, of one record for each document added to the DRIC collection and keyed on the DRIC accession number, should continue in use.
>
> That the size of this record be increased to include data relating to document movements to meet the Phase 2 requirement.
>
> That the greatest benefit would be gained, at the earliest possible time, by converting data from the bureau as follows;
>
> > First - records from 1980 to 1987. Scientific staff would be able to take advantage of the new system in a progressive manner, while still using the bureau facility.
> >
> > Second - records from 1970 to 1979.
>
> That the two sets of converted data be merged to form one large database.

At the prototyping stage DRIC was made aware that in a CAIRS system the size of a database is limited to 256Mb.

It was apparent that the extra data required for Phase 2 would cause this limit to be exceeded even for the 1980 to 1987 data.

Installation of Hardware & Software

The CAIRS software and the DEC VAX 8200 were delivered in April 1987. About the same time DEC announced that all 8200's were to be upgraded to 8250's. The computer delivered to DRIC was upgraded in October 1987 at no extra cost.

After successful acceptance trials the task of converting data from the bureau system began in May 1987.

At this stage the configuration was based on a DEC VAX 8200 computer and consisted of:

> 8 Mb of memory
> 1.8Gb of magnetic disk
> 2 Laser printers (desktop)
> 2 Dot matrix printers
> 1 Magnetic tape unit

Data Conversion

It was determined that records from the bureau database should be converted in batches of 200. At first the conversion process ran fairly smoothly, until approximately 10,000 records had been converted and the process time began to increase considerably. It was projected that the conversion of all records was now likely to take three to four times as long as originally estimated.

The conversion process was suspended and discussions took place with LFRA who explained that:

> The conversion time is directly related to the size of the database and in particular the extent to which a record is indexed.
>
> In DRIC's case almost every word in a record is indexed, except of course the normal stop words. This means the index files have many terms with several thousand postings.
>
> These factors together with the LFRA requirement that the index file must always be in absolutely precise order, which means a complete file reorganisation at each update, make the task very great indeed, since on average there are 60,000 postings per conversion batch. The end result is that the inverted file represents some 40% of the size of the searchable data. Retrieval is to a large extent governed by the size of the index files and here CAIRS benefits from its small tidy files.

As a result of these discussions various elements of the CAIRS software facilities were optimised. The conversion process was recommenced and the optimisation produced a 25% reduction in processing time.

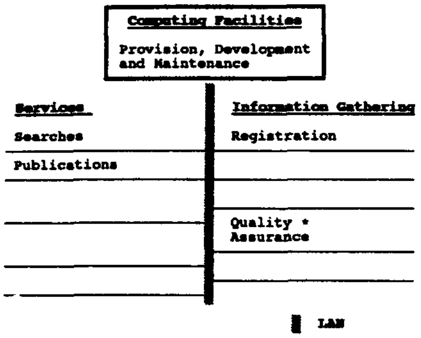
The conversion process proceeded satisfactorily and as the transfer of the 1980 to 1987 records progressed DRIC became less and less dependant on the bureau services and more able to use the new CAIRS facility.

Further problems were encountered as the conversion process handled the 1970 to 1989 data. Due to the age of the database and changes in requirement, it had been necessary to introduce changes to the data structures at the bureau. Unfortunately these changes had not been made retrospective nor had they been documented. The result was that every so often the process would fail.

In these circumstances, it took varying amounts of time to identify the fault and get LFRA to produce suitably amended software to continue the process.

Throughout the conversion process the Scientific and IT staff used the LAN with no problems or difficulties.

The functions supported by the LAN at this stage were:



* **NOTE:** When the system was accepted in May 1987 the input of new document records was transferred from the bureau to DRIC's new computing facility. This allowed DRIC to conduct Quality Assurance checks via the LAN.

3.5 Phase 2

Requirement

The Phase 2 requirement was to provide a Document Movement Control System (DMCS). This was needed to ensure safe custody of classified documents and to manage the immense task of keeping track of document movements.

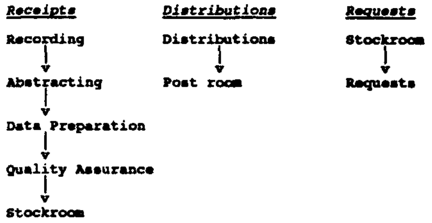
Each year the number of document movements, into, through and out of DRIC are:

> *Receipts;* 10,000 individual titles plus copies, giving a total of 20,000 to 30,000 document copies. Of the 10,000 titles received, 7,000 are added to the DRIC database. This entails 5 separate movements for each document.
>
> *Distributions;* 45,000 document copies are distributed in accordance with originator's instructions.
>
> *Requests;* 11,000 documents are supplied on request.

The total, annual document movements is in the region of 100,000. (See `Document Movement Patterns' diagram below.)

These movements were recorded manually *after* the event. This meant that it was extremely difficult to trace the exact location of an individual document.

Document Movement Patterns



Design and development

During the design and development of the DMCS there was constant consultation between the DRIC, IT section and the various user sections. Additionally, as required by national agreements, there was consultation between the IT section and Trade Union representatives regarding the introduction of new technology.

The additional terminal requirement was assessed as 27 units. Their positioning was discussed with the users and two options were considered:

> *First;* cluster the terminals at convenient locations throughout the work area. This would allow staff to use an available terminal as and when necessary.
>
> *Second;* distribute the terminals between groups of staff on the basis of 2 terminals to 3 staff.

The first option would require that;

> special areas or rooms be set aside to accommodate the terminals.
>
> additional furniture be supplied.
>
> staff would have to check the availability of a terminal for their use.
>
> staff would have to take their work, from their normal work area, to the terminal.

It was decided that the second option would provide the best environment for the staff and would cause least disruption to the normal workflow.

Bar coding was to be used to facilitate the recording of document movements, and a need for 10 bar code readers and 2 bar code printers was established.

The development of the DMCS was progressed in parallel with the procurement of the additional hardware needed.

As each software module was produced it was presented to the users for testing and acceptance.

All staff to be employed on the new DMCS attended a 3 days computer appreciation course to prepare them for the new challenge ahead and DRIC acquired a software package designed to develop and improve their keyboard skills.

Throughout the design and development of the system the IT section followed MOD approved project management and system development methodologies.

Implementation

The additional hardware was delivered in March 1988 and the DMCS went live in April 1988. During the next 6 months a variety of problems were encountered, the majority of these were easily rectified by software amendments.

However some presented greater problems. A few examples are listed below.

Terminals wrongly sited

There was an immediate spate of complaints concerning, so called, screen glare and poor lighting conditions. The real problem was that despite advice on the subject, many users had positioned their desks such that the terminal screen was in direct sunlight and they did not make use of the artificial lighting which had been installed to the approved standards for use with terminal screens.

The result was that screens reflected images *rather* than give off glare.

Unidentified functions and tasks.

Throughout the first few months previously unidentified functions and tasks were brought to light eg:

> In carrying out its duties in accordance with the DMCS requirement a section would discover a need for a specific task which *they* had not specified. On investigation it was found that, the task had previously been carried out by some other section, but had not been identified at the system development stage.
>
> There were two main reasons for this type of omission:
>
>> When DRIC moved from St Mary Cray to Glasgow in 1986 there was a complete change of staff, from top to bottom, resulting in a lack of continuity. The knowledge base was therefore too shallow to build a trouble free system.
>>
>> There was also a lack of good quality documentation of the existing manual system.

These two factors would be major impediments to the development of any system.

A knowledge base can only gain depth through time and experience and it is very difficult to recognise when sufficient depth has been attained.

Documentation of the existing system, would normally contain details of how each function should be carried out, together with workflow diagrams and examples of any forms used. Again the lack of continuity was a major factor.

To have delayed development of DMCS until a satisfactory level of knowledge was achieved would;

> have deprived DRIC of early implementation of much needed automation.
>
> not, necessarily have produced a trouble free system.

Given sound knowledge, continuity of expertise and good quality documentation of the manual system, the ***Known unknowns*** cause little problems. All that remains is to discover the ***Unknown unknowns*** and eliminate them.

Poor system response times

Complaints were made by the user about perceived poor response times. The IT section argued that the user expectation had been set too high through lack of knowledge of computing systems, and that the LAN was performing satisfactorily.

This, however was not accepted by the user who demanded that the complaint be made direct to the supplier otherwise *they (the user) would revert to a manual system.*

The truth of the matter was, not that the LAN was slow but that, when the system was used by a full user population, the screen definition and software, which were originally accepted at user trials, did not give a satisfactory level of response.

Redefinition of the screen format and amendment to the associated software soon produced a satisfactory situation.

3.6 Post Implementation Review

System Change Control (SCC) procedures, had been established in January 1988 following the completion of the Phase 1 conversion process and were used extensively throughout the Phase 2 implementation to register all problems and observations identified and record action taken to rectify them.

A Phase 2 post implementation audit, as such, was not carried out since this had effectively been on going under the SCC procedures.

A committee of representatives from the IT section and the various user sections, is responsible for the application of the SCC procedures, which are based on methodologies approved by DGITS.

This committee meets once a month and allocates priorities and sets completion dates for registered changes. It is also responsible for monitoring progress of changes in hand.

By mid 1989 the DMCS had settled down sufficiently for users to undertake the task of producing comprehensive documentation of the clerical aspects of all tasks carried out by DRIC.

Each user section was tasked with producing current working procedures. To help with this task the IT section produced modified dataflow diagrams for each known function within the DMCS. The user sections then wrote the procedures around these diagrams.

The final drafts of these procedures have been checked and approved by senior management. As changes occur the procedures are updated and reviewed. *If these are available at specification and design stage it saves valuable time later.*

3.7 Recent Activities

Having completed Phases 1 and 2 and in accordance with the original assessments, the magnetic disk capacity was increased from 1.8Gb to 3.2Gb early in 1989.

As confidence in the system grew and the user's knowledge increased, his requirements became more demanding. It soon became clear that the computer's Central Processor Unit(CPU) would not have the power to sustain the new demands being made.

A sizing exercise carried out in mid 1989 showed that the CPU usage was regularly in the region of 95%.

An acceptable usage level is in the region of 75% and so procurement action was initiated to upgrade the DEC VAX 8250 to a DEC VAX 6310. The 6310 machine offers 4 times the processing power of the original 8200 computer installed in May 1987. This is now the base of the current configuration shown at 2.2 above.

All user sections have, where necessary, been supplied with extra office furniture, including foot rests and document stands.

3.8 Observations

Those of you setting out on a similar venture may be interested in the following observations:

> Good quality documentation of the existing manual system is vital to a successful implementation.
>
> The personal needs of staff must be considered throughout development and implementation.
>
> It is most important to ensure that good relationships exist between the IT practitioners and the user at all times. Easy to say but difficult to achieve.
>
> User expectations and ambitions must not be allowed to rise beyond sensible and practical limits but how do you stop them!
>
> In DRIC's case the presence of an in-house IT section was invaluable. The use of outside consultants is fine, but when problems arise after they have gone who picks up the pieces?

Finally it must be stressed that the Technological problems were insignificant compared to the human aspects.

Gateways "Intelligents"

Christian Fluhr
Claudine Machard

Institut National des Sciences et Techniques Nucléaires
Centre d'Etudes de Saclay
91191 Gif/Yvette cedex
France

1 Sommaire:

Les bases de données documentaires accessibles en ligne sont encore utilisées par un nombre trop restreint d'utilisateurs. Ce phénomène est dû à des causes très diverses comme la difficulté d'une manipulation efficace des langages d'interrogation, l'*hétérogénéité de ces langages*, la *diversité* des interlocuteurs avec qui il faut passer contrat pour accéder à l'information, la lourdeur des procédures de connexion, les coûts, la difficulté de choisir la bonne base de données et le bon serveur pour résoudre un problème particulier.

Le problème du coût devrait être résolu avec une augmentation de l'utilisation des bases, cette augmentation passant par le traitement des autres points. La résolution des difficultés citées ci-dessus peut être réalisée en différents points de la chaîne de l'information. On peut envisager de mettre une certaine intelligence sur le poste de travail de la personne qui interroge mais cela l'oblige à disposer d'une puissance de calcul et d'une capacité de stockage importante, cela peut être, aussi, réalisé sur le serveur lui même mais cela ne facilitera l'accès qu'à ses propres bases, enfin cela peut être fait sur un système puissant autonome: le gateway, c'est probablement à ce niveau que le maximum des problèmes peut le plus facilement être résolu. C'est la raison pour laquelle, bien qu'une partie importante de ce que nous allons développer puisse être mis au compte de station ou de prétraitement sur serveur, nous avons préféré nous concentrer sur les fonctionnalités réalisables par les gateways.

Le but n'est pas ici de lister les fonctionnalités proposées par les gateways existants mais de donner l'ensemble de celles que la technologie actuelle permet d'envisager à court terme.

2 Les fonctionnalités attendues d'un gateway:

2.1 Facilité contractuelle et de comptabilité:

Le tout premier intérêt du gateway est de simplifier les aspects contractuels et comptables en ayant un seul interlocuteur pour toutes les bases auxquelles on veut accéder quel que soit le serveur. On peut même envisager une absence de contrat explicite avec un service accessible par le numéro de la carte de crédit dans les pays où cela est autorisé. L'idée française de l'accès de type kiosque, où, quel que soit le service, la facturation est incluse dans la facturation générale du téléphone, est un moyen simple de faciliter l'accès à un public nombreux. En effet, aucune démarche n'est nécessaire pour accéder à l'information au moment où elle devient nécessaire. L'inconvénient de ce service est qu'il suppose un accord avec le serveur pour une tarification particulière.

2.2 Facilité de connexion:

Là aussi, le fait de n'avoir qu'un seul interlocuteur est un atout important. La procédure de connexion est toujours identique et peut être enregistrée une fois pour toute.

2.3 Unicité du langage d'interrogation:

Une des grandes difficultés pour l'accès aux bases en ligne, même pour des documentalistes professionnels, est dûe au fait que chaque serveur a son propre langage d'interrogation. Les différents langages sont souvent très proches mais loin de faciliter la tâche cela amène à de nombreuses confusions pour ceux qui doivent changer souvent de serveur.

L'un des intérêts importants d'un gateway est d'offrir à l'utilisateur la possibilité d'un langage unique d'interrogation pour tous les serveurs. C'est le logiciel du gateway qui traduit la requête du langage du gateway vers celui du serveur. Ce type de fonctionnalité serait inutile si les serveurs adoptaient un langage commun, mais force est de constater que la tentative de généraliser l'usage du Common Command language (CCL) n'a pas été un réel succès.

La plupart des gateways permettent aussi d'utiliser le langage originel des serveurs pour les utilisateurs très habitués.

Les langages communs proposés par les gateways restent des langages qu'il faut apprendre. Le meilleur langage commun est sans doute le langage naturel, et en particulier la langue que l'on parle tous les jours même si ce n'est pas celle dans laquelle est rédigée la base de données. Nous traiterons ces points à propos des aspects "facilité" et "efficacité" de la recherche.

2.4 **Requêtes multibases et multiserveurs**:

Certains gateways offrent la possibilité de lancer en parallèle (ou au pire successivement) la même requête sur plusieurs bases même si elles sont sur des serveurs différents. Ce type de service est beaucoup plus facile à réaliser sur un gateway que sur la station de l'utilisateur. Cela implique que l'on puisse ensuite identifier les doublons de façon à présenter à l'utilisateur, dans la mesure du possible, des résultats sans redondance. Ce type de traitement peut être fait aussi bien au niveau du gateway qu'à celui du poste de l'utilisateur.

2.5 **Profil utilisateur**:

Bien servir un utilisateur, c'est être capable d'adapter son dialogue à ses connaissances. En général ce type d'adaptation est assez élémentaire, un utilisateur peut être "expert" ou "novice". On lui propose dans ce cas des dialogues plus ou moins longs pour aboutir à des commandes à exécuter. Par exemple, l'expert qui est censé connaître le langage d'interrogation, peut poser sa question directement, tandis que le novice répondra à une série de questions permettant au système de construire la requête.

Il est possible d'envisager des outils de personnalisation beaucoup plus puissants en s'inspirant des travaux récents dans le domaine de l'EIAO (Enseignement "Intelligemment" Assisté par Ordinateur)[1]. La personnalité, le domaine de connaissance, les domaines d'intérêt, la connaissance ou non du vocabulaire d'un domaine par l'usager peuvent être observés par le système d'aide au fur et à mesure du déroulement du dialogue avec le système d'interrogation. Le système d'aide qui connaît de mieux en mieux l'utilisateur peut lui proposer, alors, des aides adaptées.

2.6 **Le choix des bases**:

Les critères de choix des bases sont de deux ordres. Tout d'abord il faut trouver les bases qui sont susceptibles de répondre au problème posé. Ensuite il faut disposer de critères de choix établis sur les coûts d'interrogation ou sur la couverture des base qui peuvent être disponibles sur plusieurs serveurs.

Le problème le plus difficile est le choix des bases à partir de leur contenu. La solution la plus simple à mettre en oeuvre est de permettre un choix arborescent des domaines scientifiques qui amène l'utilisateur à situer son problème par rapport à l'ensemble des domaines possibles. Cette approche ne permet pas un choix très fin car une telle arborescence ne peut être très grande sans devenir difficile à explorer.

Une autre attitude consiste à réaliser une base des bases. Chaque article décrit le contenu d'une base sous forme de mots-clés, de vocabulaires libre ou d'un résumé. Elle contient aussi, bien entendu, la liste des serveurs qui la propose, et éventuellement le nombre de documents et la liste des champs.

Cette base peut être interrogée de manière classique en décrivant le domaine de recherche sous forme d'un fonction booléenne de mots.

Pour les systèmes disposant d'une interrogation en langage naturel avec un module de reformulation, il est possible de poser directement à la base des bases, la question que l'on va soumettre à la base qui sera choisie.

Pour cela après avoir fait une analyse linguistique de la question, le vocabulaire est normalisé, des règles de reformulation automatique de type terme spécifique vers générique sont utilisées pour généraliser la questior. et permettre un rapprochement plus facile avec les descriptions des bases qui ne peuvent être très profondes dans tous les domaines. Certains termes généraux sont produits à partir de plusieurs termes de la question confirmant qu'ils caractérisent assez bien le domaine demandé. Un calcul de proximité sémantique entre la question généralisée et les descriptions des bases est réalisé permettant de proposer des solutions dans un ordre décroissant de pertinence.

Les règles de reformulation permettant la généralisation peuvent être construites en grande partie automatiquement à partir de thesauri quand ils existent.

Exemple (prototype MERIBEL [2] s'appuyant sur SPIRIT):

Quelles sont les références sur la microanalyse d'échantillons géologiques: application à l'uranium.

Proposition de bases (par ordre décroissant de pertinence) après comparaison avec la base des bases:

EDF-DOC , INIS, INSPEC, PASCAL

Après choix d'une base et ensuite du serveur s'il y en a plusieurs, la question en langage naturel, dont l'analyse linguistique est déjà faite, peut être traduite directement en un requête adaptée au serveur.

2.7 **Facilité d'utilisation (les différents modes)**:

L'un des buts fondamentaux des gateways est de permettre une interrogation plus simple, plus conviviale et si possible plus efficace que celle dont dispose chaque serveur.

Pour faciliter l'accès à l'information pour des utilisateurs novices ou occasionnels, il peut y avoir deux approches.

La première, où l'ordinateur est maître, consiste à proposer à l'utilisateur une succession de choix pour l'amener à préciser ce qu'il veut. Ce mode répond bien au critère de non connaissance du sujet par l'utilisateur mais il peut rendre l'accès à l'information très long par un nombre très grand d'interactivité. Cette approche est la plus simple à mettre en oeuvre.

Dans l'autre approche, où l'utilisateur est maître, la question peut être directement posée. Dans la mesure où on suppose l'utilisateur novice ou occasionnel, la question la plus simple à formuler est une question en langage naturel. Dans ce cas se pose le problème de l'efficacité de la recherche ce qui signifie que tout l'effort doit être réalisé par l'ordinateur qui va devoir interpréter la question en langage naturel, l'étendre par reformulation à d'autres énoncés des mêmes thèmes, et enfin construire une stratégie de recherche dans le langage particulier du serveur choisi.

L'interrogation en langage naturel de bases de données documentaires pose des problèmes particuliers dans la mesure où l' on mélange des critères de recherche portant sur de l'information factuelle (auteur, date de parution, éditeur, etc...) et des critères de recherche textuels qui portent sur le contenu même des documents

Il convient donc de bien séparer ces deux parties car la première doit obligatoirement être traitée de manière booléenne alors que la deuxième sera traitée soit par un calcul de proximité sémantique soit par une stratégie de plusieurs questions booléennes qu'il faudra ensuite combiner.

exemple:

Un article publié par Hoppe depuis plus de 2 ans sur les interfaces d'interrogation aux bases de données documentaires.

On voit clairement ici que la première partie de la question est factuelle. "article" indique simplement la nature de l'objet cherché et n'interviendra pas ici car il s'agit d'une base d'articles. "publié par" va permettre d'identifier Hoppe comme l'auteur de l'article. Cela peut être réalisé si les relations entre les objets contenus dans la base (auteur, article, éditeurs) sont parfaitement décrites autant du point de vue sémantique que lexical ainsi que leur lien avec les champs de la base. La possibilité dans certains systèmes documentaires de demander dans quel champs un mot est présent permet de faciliter l'interprétation de la question quand la syntaxe laisse des ambiguïtés.

"publié depuis plus de 2 ans" indique que la date de publication doit être inférieure à la date du jour - 2 ans.

Le mot "sur" introduit en général le thème à rechercher mais il convient de vérifier que ce qui suit n'est pas interprétable comme critère factuel.

Le résultat de l'analyse sera donc une question booléenne portant sur des champs précis:

exemple:

AUTEUR=Huppe ET DATE<=891007

Le thème: " les interfaces d'interrogation aux bases de données documentaires" doit être traduit en langage d'interrogation du serveur. Il est souvent difficile de décomposer une question en langage naturel complexe en une seule question booléenne. En effet si l'on impose la présence de tous les mots par des "ET" on court le risque de n'avoir aucune réponse. En revanche, utiliser des "OU" donnera trop de bruit.

On peut s'appuyer sur le résultat du traitement linguistique pour introduire des opérateurs comme l'adjacence, la troncature ou le "ET".

Dans notre exemple on pourra obtenir les opérateurs suivants(ADJ signifie adjacence et ? signifie masquage d'un caractère):

interface? ADJ interrogation ET base? ADJ donnée? ADJ documentaire?

Si cela assure bien un faible bruit, le silence risque d'être considérable. Il peut être réduit par l'utilisation de termes synonymes qui sont substitués à l'aide de OU à un mot de la question ou un groupe lié par des adjacences.

exemple:

interface? ADJ interrogation ET ((base? ADJ donnée?) OU système) ADJ (documentaire? OU textuelle?)

Cette méthode aboutit à des questions qui deviennent de plus en plus complexes et on se trouve vite amené à transformer l'interrogation en une suite de questions (stratégie de recherche) en particulier dans le cas où l'interrogation porte sur des résumés ou même du texte intégral.

La méthode qui semble la plus prometteuse consiste à se rapprocher des systèmes qui proposent une interrogation pondérée fournissant une réponse hiérarchisée. La difficulté est que dans le cas de gateway l'accès au serveur ne peut se faire que par le langage d'interrogation booléen qui donne comme réponse un nombre de documents et non la liste des identificateurs de ces documents (liste inversée). Cela rend toute optimisation du processus difficile. En effet si les listes inversées de chaque mots de la question étaient disponibles, il serait possible comme dans SPIRIT, par exemple, de construire des descripteurs d'intersections de documents en commençant par les mots les plus discriminants et d'arrêter quand on est sûr d'avoir les documents les plus pertinents sans poursuivre la recherche jusqu'au bout. Il suffit ensuite de regrouper les documents par classes répondant à une question booléenne issue de la question d'origine pour présenter un résultat synthétique. Cette méthode a l'avantage de fournir des résultats équivalents à une stratégie de recherche qui poserait toute les questions booléennes que l'on peut réaliser combinatoirement à partir de la question d'origine.

exemple d'une interrogation SPIRIT [3]:

Accès à une base en texte intégral (Spécifications Techniques d'Utilisation du Minitel):

Question: effacement de l'écran de la position du curseur à la fin de l'écran

classement des documents réponses:

nro classe	nb document	intersection
1	1	effacement-écran-position-curseur,fin-ecran
2	1	effacement-écran,position-curseur
3	1	position-curseur,fin,écran
4	1	effacement,écran,position,curseur,fin

La liste est arrêtée car les critères d'optimisation considèrent que les documents qui ont des intersections plus petites et moins pondérées n'ont pas de chance de répondre à la question.

Les classes de documents sont données dans l'ordre décroissant de pertinence. Le tiret, entre deux mots, indique une relation de dépendance entre ceux-cis mais pour notre propos on peut l'assimiler à une adjacence. La virgule correspond à un ET logique. Il n'y a, bien sur, pas de OU car il s'agit d'une caractérisation de l'intersection a posteriori, dans ce cas, si deux mots sont présents, ils sont liés obligatoirement par un ET même si la présence d'un seul est nécessaire pour considérer le document comme pertinent.

Dans la mesure où l'accès aux listes inversées , n'est pas possible sur les serveurs, Il faut se contenter d'une stratégie incomplète sous peine d'avoir des temps de réponse et éventuellement des coûts très importants. On peut procéder à partir de la question la plus restrictive et la dégrader soit en passant d'un opérateur restrictif comme ADJ à un opérateur plus faible comme ET soit en retirant l'un des mots de la question.

On peut aussi comme dans les stratégies de recherche manuelles combiner deux à deux puis trois à trois les mots en se guidant sur les nombres de réponses obtenues à la suite des questions successives.

Dans tous les cas de figure on a intérêt à mettre en place une stratégie de pondération permettant de hiérarchiser des intersections entre question et documents qui contiennent des termes différents.

Cela est d'autant plus vrai qu'à partir de cette méthode permettant de considérer n'importe quel texte en langage naturel comme question, il est possible de prendre tout ou partie d'un document visualisé en réponse à une question et de proposer ce texte comme nouvelle requête. Cette technique permet de réaliser des liens hypertextes dynamiques sur des données qui n'ont pas été structurées pour un tel usage.

Une dernière approche de l'accès à l'information par l'utilisateur final est l'exploration par graphes de concepts. Cette approche peut être considérée comme intermédiaire entre une approche entièrement guidée et une approche où l'utilisateur à l'entière initiative. Elle consiste à faire naviguer l'utilisateur dans un graphe de termes liés par des relations sémantiques. Cette navigation permet à la fois à l'utilisateur d'appréhender le contenu de la base et de choisir des thèmes qui l'intéressent et qui vont composer son équation de recherche.

La manière la plus simple de réaliser une telle approche est d'utiliser le thesaurus de la base, s'il existe, comme graphe de concepts. Un exemple en est donné par Hyperline de l'ESA. Une telle approche est intéressante seulement si l'interactivité de navigation est très rapide ce qui est rarement le cas dans une dialogue avec le gateway qui se fait le plus souvent en mode caractère et à 1200 Bd. Si l'on veut qu'une telle approche soit réellement utilisée, il faut peut être envisager des architectures clients-serveur ou le logiciel client sur le poste de l'utilisateur prendra à sa charge tout le dialogue avec une représentation graphique du graphe.

3 Efficacité de l'interrogation:

Nous venons de passer en revue différents moyens de simplifier l'accès à l'information pour l'utilisateur. Mais qu'en est-il de l'efficacité de ces méthodes. Les outils linguistiques, les systèmes de reformulation que l'on peut réaliser ont atteint un niveau suffisant de fiabilité pour que leur utilisation procure non seulement un confort mais aussi un réel gain de performance au niveau de l'efficacité de la recherche. Nous allons brièvement passer en revue les caractéristiques de ces outils et permettre d'apprécier leur rôle dans le processus d'interrogation.

3.1 Les requêtes en langage naturel:

Contrairement à l'interrogation des SGBD en langage naturel qui peut s'appuyer sur une connaissance de la sémantique de la base, l'interrogation de bases documentaires se fait toujours sur des univers très larges. C'est la raison pour laquelle on fait largement appel à des niveaux d'analyse qui ne dépendent pas du domaine comme le niveau morphologique ou syntaxique et que la sémantique est limitée à de la sémantique lexicale , c'est à dire des relations sémantiques entre mots. Le seul point où une sémantique plus fine est possible est dans l'interprétation de la partie factuelle de la question et pour aider à séparer cette partie factuelle de la partie portant sur le contenu des documents.

Le traitement linguistique automatique a pour but:

- d'identifier comme le même mot des chaînes de caractères différentes (synonymes, différentes forme de sigle ou de mots composés avec tiret, formes dérivées d'un même mot)

- de lever dans la mesure du possible les homographies (même chaîne de caractères avec des significations différentes selon le contexte) par exemple ("marche" verbe ou substantif),

- de reconnaître les mots composés et plus généralement les relations de dépendance entre les termes,

- de normaliser la représentation des mots pour la recherche.

Le traitement linguistique joue aussi un rôle dans la reformulation dans la mesure où, résolvant certaines homographies, il interdit certaines inférences qui pourraient produire du bruit. (ex: poste substantif *féminin* --> P.T.T.)

Pour conclure sur l'intérêt du traitement automatique du langage naturel, il faut remarquer que, là encore, la nécessité d'interroger des données qui ont été indexées automatiquement par des systèmes ne disposant pas d'une analyse linguistique est un handicap. Les systèmes de recherche à base de linguistique donnent toute leur puissance si documents et questions sont analysés par le même traitement.

3.2 Les problèmes de reformulation:

Que la base soit indexée par un vocabulaire contrôlé ou que l'on interroge directement le résumé ou le texte intégral, un utilisateur qui exprime sa question en langage naturel a beaucoup de chance de ne pas utiliser les termes qui sont contenus dans le document. Il est donc nécessaire de produire, à partir de la formulation initiale, toutes les formulations possibles dans la langue, des mêmes concepts, de façon à retrouver tous les documents pertinents.

Si le processus de reformulation est réalisé sans précaution, cela peut avoir, l'inconvénient de produire beaucoup de bruit. Toute la difficulté d'une bonne reformulation sera de diminuer au maximum le silence sans pour autant augmenter trop le bruit. Cela est possible d'autant plus facilement que l'on fait une évaluation pondérée de l'intersection entre question et documents. Le système de pondération doit permettre, même si la reformulation produit beaucoup de documents bruyants, de les mettre en bas dans la liste des documents réponse classée par ordre décroissant de pertinence.

Les données linguistiques sur lesquelles est basée la reformulation peuvent être de plusieurs origines.

On peut distinguer des connaissances de nature générale qui peuvent servir dans n'importe quel domaine. On peut pour cela faire appel à des listes de synonymes.

Pour ce qui est de la connaissance lexicale propre à un domaine, on peut partir de thesauri existants mais qu'il faut transformer car ils contiennent souvent des informations sans intérêt pour un système qui possède un traitement linguistique: par exemple les relations entre mots dérivés d'une même racine peuvent être pris en compte automatiquement par le traitement linguistique (programme -TA-> programmeur), une relation de spécificité entre un mots et un mot composé ayant le premier mot pour tête (indexation -TS-> indexation automatique).

Il faut se rendre compte que l'utilisation des relations sémantiques comme (synonymes, termes génériques, termes spécifiques, termes associés) sont probablement trop grossières pour permettre une reformulation très fine et qu'il faudra dans l'avenir donner des relations plus proches de celles établies par les relations de dépendances (par exemple agent-action, action-objet de l'action, action-instrument d'une action, tout-partie, sorte de, etc...). Malgré cette remarque, l'usage de relations classiques en documentation donne des résultats déjà très intéressants.

Jusqu'ici nous nous sommes intéressés à l'utilisation de relations lexicales créées en dehors des bases. On peut constater qu'une partie de l'information nécessaire à la reformulation se trouve dans la base à interroger. C'est cela que l'on utilise implicitement en posant une partie d'un document pertinent comme nouvelle question (traduction de la requête initiale dans un vocabulaire plus proche de celui de la base) ou comme le fait J.C. Bassano dans le système DIALECT [4] en inférant de nouveaux mots pour la requête à partir des documents les plus pertinents fournis lors d'une première interrogation. Les mots inférés sont ceux des documents pertinents qui sont en relation de dépendance avec ceux de la question.

On peut dire enfin que des méthodes de traitement linguistique et statistique permettent, dans le cas de gros corpus, de construire automatiquement des graphes de termes significatifs reliés par des relations sémantiques. Bien que de tels traitements soient encore très imparfaits, ils permettent de diminuer considérablement le temps de production d'un thesaurus ou d'une manière générale d'un graphe de termes liés à une base.

Cela permet d'envisager une interrogation par graphes de concepts réalisés à partir d'un graphe qui reflète réellement le contenu de la base et non la nature du domaine interrogé comme ce serait le cas avec un thesaurus.

Il faut malheureusement constater que ces derniers outils supposent la disponibilité de la base en totalité pour établir ces graphes et que s'est rarement le cas de gateways sauf s'ils sont eux-mêmes serveurs.

3.3 les aspects multilingues:

Beaucoup de bases de données sont en anglais, mais d' autres existent en japonais, en anglais, en allemand, en espagnol, etc... La possibilité de comprendre un document dans une langue étrangère n'implique pas forcément que l'on maîtrise suffisamment celle-ci pour interroger efficacement. La possibilité d'interroger dans sa langue maternelle des bases exprimées dans d'autres langues semble d'un grand intérêt.

Cette interrogation multilingue présente même un intérêt au cas où l'utilisateur ne comprend pas du tout la langue de la base. Le fait de trouver des documents qui semblent pertinents avec éventuellement un contrôle par une traduction automatique, même élémentaire, permet une prise de décision de traduction avec un minimum de risque.

Ce problème de l'interrogation multilingue est étudié dans le cadre du projet ESPRIT EMIR [5]. Ce projet porte sur l'interrogation de bases de données textuelles mais il pourrait s'adapter à l'introduction d'une interrogation multilingue dans un gateway.

L'interrogation multilingue est prise comme un problème particulier de reformulation. Contrairement à ce qui se passerait dans le cas de traduction automatique de la question, l'interrogation multilingue ne risque pas de provoquer du silence et avec un bon système de pondération le bruit ne doit pas être trop nuisible.

En effet en cas de traduction automatique de la question, le système est obligé de choisir une seule traduction par mot. si le système de TAO fait un contresens, le système de recherche partira dans une mauvaise direction.

Au contraire, la reformulation multilingue essaiera de rechercher toutes les traductions possibles. Cela pourrait provoquer beaucoup de bruit, mais associé à une évaluation pondérée de la reformulation, les expériences montrent que l'on peut voir apparaître en tête de liste les documents les plus pertinents avec aussi, qui plus est, la bonne traduction des mots de la question. Cela montre qu'une base de données textuelle dans un domaine peut servir de base de connaissances pour choisir un bonne traduction d'un mot ambigu en cas de traduction automatique.

Exemple d'interrogation multilingue (sur une base en texte intégral de réglementation nucléaire):

Question:

management of nuclear wastes

règles de reformulation multilingue utilisée sur les unitermes:

management--->maniement/direction /conduite/gérance/gestion/ /exploitation/adresse/savoir-faire/administration/direction

nuclear ---> nucléaire

waste --->gaspillage/déperdition/ détérioration/dépérissement/freinte/déchet/débris/résidu/rebut/déblais

Classement des documents réponses par ordres décroissant de pertinence:

nro classe	nb document
intersection	
1	1

gestion-déchets, nucléaire, résidus,débris, administration

2 3
gestion-déchets,nucléaire,
administration

3 5
gestion-déchets, nucléaire

Pour traiter un tel exemple il faut disposer d'une analyse syntaxique qui permet:

- de conditionner la traduction au rôle syntaxique du mot ("arrêté" participe passé --> "stopped")("arrêté" substantif-->"decree")

- de faire les transformations syntaxiques en particulier dans le groupe nominal (nuclear waste --> déchet nucléaire)

- de traduire globalement des expressions (pomme de terre --> potato)

- de permettre, par une reformulation en langue source, une meilleure efficacité pour trouver la bonne traduction.

- permettre de diminuer encore le silence par une reformulation en langue cible.

Il faut souligner que le point délicat est la traduction des expressions. Certaines expressions, en nombre limité, considérées comme expressions idiomatiques peuvent être repérées, introduites dans un dictionnaire , et donc être reconnues lors de l'analyse de la question en langue source. Etant donné le nombre de ces expressions, celles-ci peuvent aussi être introduites dans les règles de transfert langue source-langue cible.

Mais en ce qui concerne les autres expressions, elles peuvent souvent être traduites mot à mot mais ceci demande une combinatoire de recherche dans le lexique de la base ce qui peut être coûteux. Il existe aussi malheureusement beaucoup d'exemples qui ne peuvent être traduits mot à mot (air bag --> sac gonflable).

Il est très coûteux de construire de vastes lexiques d'expressions avec leur traduction faite entièrement à la main. Une solution qui va être expérimentée dans le cadre du projet EMIR est la construction de dictionnaires de transfert d'expressions à partir de textes déjà traduits.

Si on ne dispose pas de tels textes traduits, il reste la ressource de traiter des textes monolingues pour repérer les expressions significatives du domaine grâce à un traitement linguistique et statistique, et ensuite d'en faire une traduction manuelle.

4 Quelques exemples de systèmes existants ou projets en cours de réalisation:

Les fonctionnalités décrites plus hauts sont celles que permet la technologie actuelle. On peut donc envisager des gateways dans un avenir proche qui comprendraient tout ou partie de ces fonctionnalités.

Les gateways actuellement en service comme Easynet [6][7], Infotap, l'ESA, proposent les services de base comme le contrat unique quel que soit le serveur, un langage commun d'interrogation et le choix de la base.

L'ESA propose Hyperline qui est le tout début de ce qui pourrait être une interrogation par graphes de concepts.

Certains projets s'attaquent au problème de l'efficacité de l'interrogation plus en profondeur et y incluent souvent une part de traitement linguistique.

On peut citer par exemple:

- Le projet IMPACT "MITI" qui va permettre une interrogation multilingue de sept serveurs européens en anglais, français, allemand, espagnol dans le domaine de la technologie et de l'environnement. Ce travail s'appuie sur les recherches menées auparavant par les partenaires (système PLEXUS B. Vickery Tome Ass.[8], EURISKO Barthes et Glize IRIT Toulouse [9], EXPRESS Ulrich Hoppe GMD-IPSI *Darmstadt*[10], *Primus*, LC-TOP et les dictionnaires électroniques de Softex).

- Le projet IMPACT "CARTINFO" qui va permettre d'ouvrir des services ciblés sur les PME-PMI dans différents pays de la Communauté. Une étude de marché a permis de déterminer les besoins précis de ces entreprises et le système va pouvoir répondre à des requêtes préidentifiées. L'accent a été mis aussi sur la distribution rapide de l'information, résultat de la recherche, par messagerie électronique, fax ou courrier.

- l'anté-serveur [11] TRIEL réalisé en collaboration entre TRIEL et l'Université de Caen. Ce projet s'attaque aux problèmes linguistiques (syntaxiques et sémantique) pour le choix des bases et la traduction des questions.

5 conclusion:

L'élargissement du nombre d'utilisateurs interrogeant les bases de données en ligne est indispensable pour que cette activité acquiert un caractère de rentabilité réel. Cela ne peut se faire que si l'accès est facile, efficace, et peu coûteux.

Les gateways ont un rôle important à jouer pour répondre à cette demande. Bien que la résolution de certains des problèmes puisse être implantée ailleurs (sur le poste de l'utilisateur ou sur le serveur) , le gateway est le seul a pouvoir simplifier l'aspect contractuel en proposant un seul accord pour un service diversifié. L'accès multiserveur en parallèle est aussi un service que seul le gateway peut proposer bien que l'on puisse envisager, strictement, un poste d'utilisateur doté d'un capacité de liaison multiple.

6 Bibliographie:

[1] Claes G., Ounis O., Razoarivelo Z., Salembier P., Shridaran M.S., "STARGUIDE: a generator for self tutorials", RIAO88, Cambridge, march 1988, editor CID 36 bis rue Ballu Paris.

[2] Machard C., "station "intelligente" d'interogation de bases de données publiques en ligne", mémoire d'ingénieur CNAM, Paris , février 1991.

[3] Debili F., Fluhr C., Radasoa P., "About reformulation in full text IRS", Information processing and management, VOL 25, No 6, pp 647-657, 1989.

[4] Bassano J.-C., "DIALECT: an expert assistant for information retrieval", Canadian AI conference, Montreal, May 1986.

[5] Fluhr C.,"Multilingual Information", Pacific Rim International Conference on Artificial Intelligence (PRICAI), IAIS 90, "AI and Large-Scale Information", Nagoya, 14-16 November 1990.

[6] Larsen G., Villumsen S., "Intelligent Gateways: Evaluation of Easynet - an end-user test.", Congres Online 86, Londres, december 1986.

[7] O'Leary M., "Easynet revisited, pushing the Online frontier", Online, p 269-290, 5-6 December 1988.

[8] Vickery A., Brooks H.M.,"Plexus: "the expert system for referral", Information Processing and management, VOL 23, No 2, pp 99-117, 1987.

[9] Barthes C., Glize J-P, "Planning an expert system for automated information retrieval EURISKO", 11th International Conference on research and Development in Information retrieval, ACM Grenoble, 13-15 june 1988.

[10] Hoppe H. Ulrich, "Intelligent Access to Public data banks, Pacific Rim International Conference on Artificial Intelligence (PRICAI), IAIS 90, "AI and Large-Scale Information", Nagoya, 14-16 November 1990.

[11] Schiavo J., Interrogation libre d'un système documentaire: l'antéserveur Triel, Conférence Génie Linguistique 91, vol. 1, Versailles, 16-17 Janvier 1991.

[12] Efthimiadis Efthimis N., "Progress in documentation. Online searching aids: a review of front ends, gateways and other interfaces, Journal of documentation, vol. 46 n° 3, september 1990.

[13] Ermine J-L, Le Blanc B., "Rhéthorique et reconnaissance des formes pour l'ajustement de requêtes documentaires", Congrès RFIA'91, Lyon, Novembre 1991.

[14] Hawkins D.T., "Applications of artificial Intelligence (AI) an expert system for online searching, Online, p 31-42, january 1988.

[15] Brian C. Vickery, "Intelligent interfaces for user-friendly access to databases and electronic information services", state of the art survey, CEC report contract ML-60, Luxembourg, june 1989.

UTILISATION DES SYSTEMES EXPERTS DANS DES INTERFACES POUR LA RECHERCHE DOCUMENTAIRE

(Use of Expert Systems as User Interfaces in Information Retrieval)

par

J-C. Bassano, D. Archambault, G. Desrocques et A. Mekaouche
Lifo Université d'Orleans
BP 6729
45067 Orleans, Cedex 2
France

Par les interfaces présentés ici, on ne cherche pas à fournir directement à l'utilisateur la réponse à la question posée. On lui propose un texte ou un ensemble de textes. Cela se traduit par des interfaces qui tentent de mettre à sa disposition toute une panoplie d'outils. Il s'agit d'abord des moyens d'une interactivité élaborée et conviviale. Il s'agit ensuite de techniques intelligentes et efficaces d'appariement progressif du "sens" entre la question et l'ensemble des documents.

Pour faire coopérer des outils dont l'efficacité est - pour la plupart - déjà éprouvé, on constate que l'on est récemment passé d'interfaces monolithiques à des architectures hybrides ou "multi-experts". L'évolution des recherches se dirige maintenant vers des réalisations "connexionnistes".

1. Interfaces conviviales ou intelligentes: des constructions monolithiques aux architectures hybrides multi-experts

1.1. Position du problème.

Notre projet est d'observer certaines procédures complexes impliquées dans des interfaces sur des systèmes de recherche documentaire. Ces procédures contribuent à donner une apparence conviviale, puissante et "intelligente" lors de l'utilisation du système. Il faut d'abord souligner que des approches apparemment différentes, notamment la plupart des discours sur des interfaces conviviales en "langage naturel" ou sur des interfaces intelligentes s'appuyant sur les techniques de "l'intelligence artificielle", relèvent également pour la plupart du domaine abordé ici. On observe désormais un nombre important d'applications utilisant des mécanismes de type "système expert" lors de la mise en place d'interfaces performantes dans le domaine de la recherche documentaire. Paice et Smith [Paice C. 86; Smith 86] ont établi un premier état de l'art sur ces thèmes. Fox, Vickery et Dachelet ont, pour leur part, relevé dans des études bibliographiques des travaux intéressants [Fox 87; Vickery 89; Dachelet 90]. Nous centrons cet état de l'art non seulement sur l'identification d'interfaces de type "systèmes experts" dans des procédures de recherche documentaire, mais également sur une présentation de l'architecture utilisée dans ces interfaces.

L'analyse et la compréhension du langage naturel occupent habituellement une position centrale dans toutes ces interfaces, que l'on s'intéresse simplement à l'analyse de la question de l'utilisateur ou qu'il s'agisse - de manière plus ambitieuse - d'une première étape dans la compréhension de documents textuels enregistrés dans les bases. Les techniques de représentation du sens ont notamment évolué d'une approche essentiellement statistique vers des approches linguistico-conceptuelles. D'une approche linguistique, on est ainsi passé insensiblement à une approche relevant plus nettement de l'intelligence artificielle: des bases de connaissances interviennent, elles sont constituées de concepts entre lesquels divers types de relations sans statut linguistique sont établis. Enfin, on se préoccupe de plus en plus d'améliorer le fonctionnement du système et de satisfaire les besoins d'information d'un utilisateur particulier en tenant compte de ses caractéristiques propres lors d'une recherche. On tire en cela parti d'une approche "recherche cognitive".

On cherche habituellement, par des systèmes dits "experts" s'appuyant explicitement sur des règles, à résoudre efficacement des problèmes complexes dans des univers sémantiques restreints. On dispose pour cela de grandes quantités de connaissances sur des sujets précis. Cette expertise du domaine, correspondant à une compétence et à un savoir-faire acquis par des spécialistes, doit pouvoir être acceptée et communiquée sous une forme souple et déclarative de règles. Un mécanime inférentiel exploite dynamiquement et au mieux, dans chaque cas particulier, ces bases de connaissances internes généralement exprimées sous forme de règles. Ainsi, un système expert tend à captur:ᵣ les connaissances d'un ou plusieurs experts agissant dans des domaines spécialisés. Dans notre cas, il s'agit **d'abord de l'intermédiaire humain** qu'est le **documentaliste**. Celui-ci est avant tout un généraliste qui utilise ses compétences et son expérience pour guider et aider le demandeur d'information. Aussi, l'expertise qu'il s'agit d'introduire dans le système, est-elle d'abord centrée sur la connaissance des outils et des techniques nécessaires pour localiser et choisir une base de données, puis sur la maîtrise des procédures permettant de manipuler les informations enregistrées dans la base choisie, enfin sur les capacités pour comprendre la requête de l'utilisateur à partir de connaissances linguistiques générales et pour la formuler de manière adaptée.

Mais il peut s'agir également d'une **analyse plus approfondie du sens** des documents et des questions. Le fonds documentaire couvert par les bases de données auxquelles s'adresse l'utilisateur est très large. En raison de la nature encyclopédique du domaine concerné, une recherche sur le contenu ainsi que la transformation des termes de la requête initiale demeurent un problème très difficile. Ce problème peut à nouveau être abordé à partir de connaissances linguistiques générales portant notamment sur la paraphrase. Il peut l'être également à partir des connaissances d'un spécialiste du domaine traité. Le problème est alors partiellement résolu par l'élaboration et l'utilisation de bases de connaissances spécifiques (de thesauri, de réseaux sémantiques, etc.) qui complètent les connaissances générales du documentaliste sur la langue.

On met ainsi en évidence, au niveau du **documentaliste** comme de l'**analyste**, des sous- problèmes parfaitement définis et correspondant à des traitements spécifiques. On devine également que les étapes d'inférences autonomes sous le contrôle du système, resteront toujours moins nombreuses que dans les systèmes experts plus traditionnels. Les règles seront généralement nombreuses et parfois mal définies, elles seront peu consistantes et souvent redondantes. Il est donc nécessaire de les structurer et de disposer de méta-règles ou de règles stratégiques. Aussi, derrière les avancées théoriques de ces dernières années, on remarque que l'on est passé de tentatives d'implantation d'interfaces souvent monolithiques à des implémentations de plus en plus complexes et hybrides prenant appui sur des architectures dites de **systèmes multi-experts**. Cette approche technique apparaît donc comme une voie efficace pour résoudre des problèmes posés par la recherche documentaire. Des modules **spécialistes** - dont l'activité est essentiellement fondée sur des opérations de "filtrage", de choix et de propagation de connaissances très spécifiques- sont eux aussi pilotés par des modules experts en **stratégies**. Ces **stratèges**, disposent dans leur propre base de connaissances constituée de règles stratégiques et heuristiques. Ils guident le fonctionnement des spécialistes en déclenchant, en organisant et en coordonnant leur travail.

On introduit ainsi une différence entre des interfaces **conviviales** rendant essentiellement compte des actions du documentaliste et des interfaces **intelligentes** traduisant des capacités de compréhension approfondie de l'analyste. Nous situerons ces deux niveaux - d'une complexité croissante quant à l'architecture et la puissance des systèmes multi-experts impliqués - dans les sous-sections suivantes 1.2 et 1.3. Nous en reprendrons les aspects essentiels à travers la présentation d'applications dans les paragraphes 2 et 3.

1.2. Les interfaces dites évoluées ou conviviales, interfaces relativement simples disposant généralement de possibilités de dialogue en langage libre

Ces interfaces regroupent des outils que l'on peut ajouter à un système documentaire - ou éventuellement à un système de gestion de bases de données utilisé pour une application documentaire - pour obtenir un environnement convivial qui en facilite l'utilisation. La recherche porte alors sur l'ergonomie des interfaces. L'objectif est de rendre l'interaction de l'utilisateur avec le système documentaire plus souple et plus agréable. Les interfaces graphiques, les interfaces en langage naturel répondent à ce souci. Souvent ces interfaces sont mises en oeuvre dans un contexte multi-bases. L'utilisateur dispose alors d'un ensemble plus ou moins complet de procédures spécifiques complémentaires. Ces procédures peuvent être réalisées sous la forme de spécialistes ou d'experts qui assistent l'utilisateur dans le choix des bases et qui prennent en charge les procédures de connexion et de communication à distance [voir par exemple MESSIDOR de Molinoux, SCI-MATE de Stout, IN-SEARCH, SMARD de Sellami, MYRIADES de Bernard, EURISKO de Barthes;...]. Dans ces interfaces, on développe tout spécialement la convivialité des processus d'échange par un premier traitement élémentaire du langage de l'utilisateur. On prend en charge la gestion du dialogue. On aide l'utilisateur à respecter les règles syntaxiques et les règles de transcodification vers des langages d'interrogation spécifiques. Il s'agit finalement d'interfaces évoluées qui sont cependant relativement mécaniques.

Aussi, les interfaces reprises plus loin dans le paragraphe 2 sont celles qui permettent avant tout une **interrogation en langue naturelle**. Dans certains cas les requêtes peuvent porter sur le contenu de zones de textes enregistrées dans la base. Ces traitements linguistiques restent relativement simples dans le contexte de la recherche documentaire: analyse de la structure de la requête et mise en évidence de la nature des informations demandées (données signalétiques et/ou recherche sur le contenu); appariement autour des formes identifiées dans cette requête initiale pour une recherche sur le contenu. Pourtant cette interface d'interrogation en **langage libre** nécessite déjà à elle seule l'intervention de techniques liées aux systèmes multi-experts. De tels systèmes d'analyse et de compréhension du langage naturel sont maintenant eux-même conçus autour d'une architecture dite "multi-experts" et utilisent généralement des outils de communication fondés sur la technique du "tableau noir" ou sur l'échange de messages.

1.3. Les interfaces intelligentes fondées sur des représentations élaborées des connaissances du domaine (la base de données est, ou est complétée par, une base de connaissances).

Une analyse plus approfondie du sens des documents et des questions est nécessaire pour permettre un meilleur appariement. Des systèmes de recherche documentaire comportent, au delà des phases d'analyse linguistique déjà évoquées en 1.2, des phases complémentaires tendant à élaborer une représentation sémantique et/ou pragmatique des énoncés. Ces procédures d'analyseconduisent à une structuration de la collection de textes.

Une **première approche** construit et exploite alors de manière systématique des représentations sémantiques et pragmatiques issues du contenu informationnel de la base elle-même. La base de textes peut être considérée comme une base de connaissances assertionnelle. C'est généralement la fonctionnalité de **paraphrasage** qui est surtout invoquée. Les paraphrases générées par transformation lexicales et/ou linguistique démultiplient les possibilités d'appariement formel entre requêtes et documents. S'appuyant sur des connaissances essentiellement linguistiques, donc indépendantes du domaine d'application, les transformations évoquées restent générales. Ces techniques de paraphrasage manipulent des représentations assez proches de la surface langagière et ne nécessitent éventuellement pas de recourir à un niveau de représentation des connaissances d'ordre conceptuel plus profond.

Mais, **une deuxième approche** se fonde directement sur une notion de "méta-documents" qui sont constitués, directement et souvent a priori, sans passer par une expression textuelle. Ces systèmes s'appuient ainsi sur l'utilisation de "**bases de connaissances**" très élaborées. Dans ce cas la base documentaire traditionnelle - ou assertionnelle - est partiellement reléguée à l'arrière plan. On privilégie alors des "bases de connaissances terminologiques" relatives aux domaines couverts. Il en est ainsi du thésaurus, outil forgé pour la représentation globale de la base. Si la traduction des phrases d'un texte en énoncés est une chose difficile, l'extraction automatique de la représentation d'une phrase dans les termes d'une "base de connaissances" l'est tout autant. On peut, dans certains cas intermédiaires et à mi-chemin entre les deux approches, tenter de décrire les "méta-documents" en passant par une expression textuelle... On peut également se poser le problème de constituer cette base de connaissances à partir d'encyclopédies déjà existantes.

Dans ces deux approches, des adaptations des mécanismes habituels de l'intelligence artificielle et des systèmes experts doivent être réalisées en vue de leur utilisation pour des bases documentaires vastes et de nature encyclopédique: jusqu'ici, les textes que l'intelligence artificielle a réussi à traiter sont courts et relèvent de domaines restreints. Aussi, il faut également faire apparaître les contraintes propres aux applications documentaires abordant - de manière souvent atypique - des domaines de connaissances très larges. Très souvent, par un processus dynamique et une stratégie naturelle de recherche progressive de l'information, les résultats d'une première requête sont utilisés automatiquement pour reformuler la demande.

Cette classification a fait apparaître des interventions de systèmes (multi) experts de plus en plus complètes et imbriquées. Les systèmes fondés sur des "bases de connaissances" utilisent probablement des interfaces conviviales et effectuent nécessairement des analyses linguistiques des documents (en plus de l'analyse de la question). Les modèles de représentation des connaissances sont étroitement liés à la compréhension et au traitement du langage naturel. Aussi, tous les systèmes cherchant à apparier le sens disposent généralement et également d'interfaces en langage naturel...Des systèmes seront donc cités plusieurs fois en exemple en fonction des aspects précis auxquels on s'intéresse: l'utilisateur peut effectivement déclencher dans ces systèmes des experts de plus en plus complexes.

2. Interfaces conviviales disposant notamment de possibilités d'interrogation en langage naturel: une première étape vers des systèmes multi-experts.

2.1. Un élément de référence: l'interrogation en langage naturel de systèmes de gestion de bases de données.

Ces systèmes permettent aux utilisateurs d'interroger à partir d'une requête en langage naturel des bases de données contenant des informations structurées.

Les systèmes de gestion de bases de données manipulent essentiellement des données structurées. Certains peuvent cependant permettre l'enregistrement de chaînes de caractères et disposent de primitives autorisant des manipulations spécifiques de textes et de documents. On peut ainsi citer [Croft et alii. 1982; McLeod et Crawford 1983; Miranda 1983; Bancilhon et Richard 1987]. Autour de tels systèmes ont peut déjà proposer des interfaces en langue naturel effectuant des sélections sur les différents champs et comportant des fonctions de visualisation des textes et des documents. Les systèmes de Normier, Hendrix, Bobrow [Normier et alii. 1985; Hendrix 1982; Bobrow et Bates 1983] montrent d'une manière générale les techniques utilisées pour la mise en place d'interfaces en langage naturel pour des systèmes classiques de gestion de bases de données. Les travaux sur Saphir et l'extension du système Smart-relationnel proposée par Fox illustrent plus particulièrement l'évolution vers des analyseurs à base de règles [Normier B. et alii. 1985; Fox 1981].

Lorsque l'on soumet à SAPHIR la requête "**professeurs de maths célibataires qui enseignent dans plusieurs établissements**", il l'analyse comme "**je vais vous donner les noms des établissements et des professeurs, professeurs dans la matière mathématique et dont la situation de famille est *célibataire et qui ont un poste dans plus d'un* établissement**" sous une forme qui paraphrase la requête QBE ou SQL sous-jacente

Même si certains champs de la base contiennent des zones de texte, on ne dispose généralement pas des procédures de manipulation des termes des textes. Or, ces procédures sont nécessaire à la recherche documentaire.

2.2. Adaptation à la recherche documentaire

D'autres systèmes permettent déjà, au delà de simples manipulations de textes, une recherche sur le contenu. Cette recherche correspond à l'appariement de termes dits "descripteurs". La demande d'information, formulée initialement en langue naturelle, aboutit généralement à la recherche des solutions d'une équation booléenne ou "booléenne étendue" de descripteurs. Dans cette forme élémentaire d'accès aux documents par leur contenu, la représentation du "sens" des documents textuels est liée au problème de l'indexation automatique. Des procédures d'indexation automatiques, plus ou moins puissantes, sont donc également ajoutées pour ramener les textes de la base à un ensemble de descripteurs. DIALECT1 [Bassano 1986] a été construit comme une interface sur le système classique de gestion de base de données ADABAS. Ce progiciel possède des caractéristiques appropriées pour la mise en place d'une telle interface. Lors de l'insertion des textes dans la base, on construit un fichier inverse. Les mots pleins sont identifiés et les mots grammaticaux trop fréquents et sans poids sémantique sont retirés. Il s'agit donc essentiellement d'une reconnaissance de la forme des mots à partir de délimiteurs et de signes de ponctuation, de l'élimination des séquences terminales pour réaliser une normalisation des formes.

Par exemple, pour la requête "**recherche des documents traitant des systèmes documentaires, écrits par Bassano**". DIALECT: identifie les termes "système" et "documentaire" pour le champ descripteur et les relie par "ET", relève le terme "Bassano" pour le champ auteur. En cas d'echec, le système transformera certains opérateurs "ET" en "OU".

Une nouvelle étude est en cours autour du logiciel orienté objets O2 [projet Altair, Bancilhon et alii. 1987].

Mais on peut également utiliser des systèmes construits dès l'origine spécifiquement pour la recherche documentaire. Ces systèmes prévoient donc déjà des possibilités de recherche sur le contenu.

Le système IOTA de l'équipe de Chiaramella [Defude 1984, etc.] propose une interface en langage naturel pour un système de recherche bibliographique. L'analyse de la requête permet le passage d'une forme syntaxiquement riche et donc potentiellement ambiguë à la forme simple et non ambiguë de l'équation de recherche booléenne. L'analyse est centrée sur la reconnaissance des groupes nominaux, les mots employés dans la requête sont connus du système. On identifie les différents concepts et les opérateurs logiques les reliant, on associe à chaque concept les termes correspondant du langage d'indexation. Le système ALEXIS de la société Erli ou le système de Gauch [Gauch 1988] sont d'autres exemples de ce type d'interface sur des systèmes documentaires spécifiques.

A partir de la question "**mobilier contemporain en bois massif**" ALEXIS, DIALECT ou IOTA *identifient dans un premier temps les syntagmes* "**mobilier contemporain**" et "**meuble en bois massif**".

Le système SPIRIT utilise, pour l'indexation automatique, des outils syntaxiques et lexicaux puissants lui permettant également d'identifier des unités syntagmatiques valides. Lors de la recherche portant sur des données textuelles, il met en oeuvre une procédure d'appariement complexe. Le principe est d'une part, de conférer aux termes des poids à partir de techniques probabilistes et d'autre part de calculer des distances entre requêtes et documents.

A partir de la question "**un collier en dent de requin**", SPIRIT identifie les mots vides (un, en, de) et les mots pleins (collier, dent, requin). Il repère également le syntagme "dent-requin". La classe des documents les plus pertinents contient des textes possédant les termes "**dent-requin**" et "**collier**". La suivante ne contiendra que des documents repérés par le syntagme "**dent-reqin**". Les documents de la troisième ne posséderont plus que le terme "**collier**", etc.

On propose une nouvelle version de DIALECT interfacée sur le système Spirit [DIALECT2 91 de Mekaouche].

L'analyse de la requête initiale, les techniques linguistiques utilisées pour l'indexation automatique des textes se fondent donc sur des traitements morphologiques et syntaxiques relativement simples. Ces traitement permettent d'extraire des unités syntagmatiques *syntaxiquement* (et *sémantiquement* ?) valides. Une grande part des travaux du CRISS de Rouault et de l'équipe de Bouché est également consacrée à une entreprise de ce type. On peut également consulter les travaux de Lancel. Des orientations vers une architecture multi-experts apparaissent très clairement dans tous ces travaux. Une *attention toute particulière doit être portée* aux analyseurs à base de règles [Rouault et Lallich; Marcus; Charniack; Rady; Fouquéré; etc.] qui sont - ou qui peuvent être - utilisés dans ces interfaces. Différents spécialistes manipulent des connaissances linguistiques spécifiques: spécialiste des entrées lexicales, spécialiste des homographes, spécialistes des lexis, spécialiste des mots gammaticaux par exemple [Mekaouche 90]. A ces experts linguistes s'ajoutent éventuellement les experts déjà cités en 1.2 sur le choix des bases et sur les techniques de communication.

3. Interfaces intelligentes: accès à l'information lorsque la base de données est, ou est complétée par, une base de connaissances.

3.1. utilisation du paraphrasage et des textes de la base comme énoncés d'une composante assertionnelle.

3.1.1. Un élément de référence: des systèmes classiques du type question-réponse prenant appui sur des bases de textes (avec ou sans activité inférentielle).

Parmi les approches strictement linguistiques, mentionnons celle pratiquée depuis près de vingt ans par l'équipe de N. Sager [Sager N. 78 et le "linguistic string project]. Les textes sont des textes scientifiques et techniques appartenant à un domaine spécifique homogène. Les traitements linguistiques aboutissent à pouvoir convertir les énoncés en instances d'un nombre réduit de schémas d'énoncés. Ceux-ci donnent lieu à une formalisation sous forme de tables relationnelles. Ces tables sont utilisées ensuite pour différentes applications question-réponses reposant finalement sur une forme "d'interrogation de bases de données structurées". La réussite attestée de la procédure d'analyse est liée à l'homogénéité sémantique des classes d'équivalence syntaxique et à la récurrence d'un nombre faible de schémas canoniques d'énoncés. Lors de la formulation de la requête, on n'observe pas d'activité inférentielle.

Unc autre approche, encore nettement marquée par la linguistique mais utilisant parfois des règles ou des représentations de nature sémantique, peut être citée ici. Elle procède essentiellement à des analyses lexicales morphologiques et syntaxiques. La représentation syntaxique permet éventuellement de dériver une représentation sémantique. Les systèmes de D. Kayser et de son équipe en sont un bon exemple. Ils travaillent sur le texte de la question et sur des fragments de textes trouvés dans la base de connaissances très spécialisée pour produire, par des inférences de niveaux variables, la réponse à la question. Dans certains cas, on reste très proche de la structure de surface des textes. On dispose clairement de mécanismes de reformulation et d'inférence.

A la question **"est-il normal qu'un enfant de 6 mois ne sache pas encore marcher?"**, la réponse est "oui, avec une plausibilité de 90% car nous avons trouvé dans la base QUID l'affirmation: **l'enfant commence à marcher seul (à l'age) de 12 à 18 mois"** Cette déduction a utilisé un raisonnement bidirectionnel et a créé 17 noeuds. Neuf règles ont été successivement déclenchées.

Dans le paragraphe suivant, nous montrons que des adaptation de ces techniques permettent d'obtenir des résultats intéressants lorsque les textes sont longs, nombreux et portent sur des domaines non spécifiques. De façon contrôlée, des fragments de textes résultant d'une première recherche sont utilisés pour reformuler la requête. L'augmentation de la précision des résultats est souvent spectaculaire [voir par exemple Salton G. et Bassano J-C.].

3.1.2. Adaptation à la recherche documentaire.

Le système SPIRIT de C. Fluhr et de son équipe visualise un ensemble de textes susceptibles de répondre à une requête. ces textes sont classés par ordre décroissant de pertinence. Le système permet d'utiliser le texte le plus pertinent comme nouvelle question à soumettre au système. Il s'agit alors d'une procédure avec reformulation par le document le plus pertinent. Dans SPIRIT [Fluhr 85 et 91, Debili 88], c'est donc l'utilisateur qui sélectionne, parmi les textes retrouvés les parties qui peuvent servir de nouveaux points de départ. On établit ainsi une liaison dynamique entre des textes ou des parties de textes, liaisons très voisines de la notion d'hypertexte dynamique. Mais l'on n'observe pas de tentative de reformulation automatique, progressive et autonome de la part du système.

Le système DIALECT utilise également les résultats des recherches précédentes pour reformuler la requête. Il ne retient que les phrases les plus pertinentes et non une partie du document. DIALECT conduit cette opération automatiquement en mettant en oeuvre des outils d'analyse linguistique et des procédures de contrôle. La question de l'utilisateur, écrite en langage naturel, est analysée puis utilisée pour extraire un premier noyau de "zones de texte" très pertinentes. Ces zones sont à leur tour analysées et exploitées en vue d'enrichir la question. Le système s'appuie essentiellement sur des procédures d'analyse distributionnelle permettant de repérer des régularités syntaxiques formelles. Ces reformulations sont relancées automatiquement jusqu'à l'obtention d'une condition d'arrêt.

Partant d'une requête sur **"l'évaluation des systèmes de documentation"**, la reformulation propose une requête transformée que l'on peut paraphraser par **"modèle ou critère pour une évaluation du coût, de l'efficacité ou des performances de systèmes ou programmes; ces systèmes ou programmes concernent la documentation automatisée, la recherche documentaire ou bibliograhique on-line; (il peut s'agir de) méthodes traditionnelles d'évaluation"**.

De l'approche linguistique, nous passons insensiblement à l'approche intelligence artificielle en privilégiant la représentation et la modélisation des connaissances.

3.2. Utilisation d'une composante terminologique.

3.2.1. Un élément de référence: les systèmes question-réponse classiques fondés sur une représentation élaborée des connaissances

Ces systèmes utilisent de larges bases de connaissances construites autour de domaines spécifiques. Des formalismes du type grammaires de cas, scénarios (frame), réseaux sémantiques, scripts, etc. sont utilisés pour représenter les connaisssances. On peut citer ici les systèmes de l'équipe de Schank ou ceux de l'équipe de G. Sabah [Vilnat 84].

Un système prototype a été proposé par Vilnat pour l'interrogation des "pages jaunes" de l'annuaire électronique. Partant de la question **"je cherche à réparer mon auto-radio"**, ce système explique qu'il vaut mieux passer par un garagiste puisque l'auto-radio fait partie de l'automobile. Après s'être renseigné sur l'adresse du demandeur, il propose les garagistes les plus proches.

Pour déplacer le problème lié à la construction des ces bases de connaissances, certains travaux proposent l'utilisation du langage naturel, ou plutôt d'un langage pseudo-naturel, pour exprimer ces connaissances [Wilks, Zarri par exemple]. D'autres cherchent à les construire automatiquement à partir d'encyclopédies. Ces techniques restent très largement du domaine de la spéculation.

Là encore, des adaptations intéressantes pour la recherche documentaire sont présentées dans le paragraphe suivant.

3.2.2. Adaptation à la recherche documentaire.

Au lieu d'essayer d'approfondir l'analyse linguistique d'un texte, le système RUBRIC de Tong recherche des indices - sous forme de mots - lui permettant de repérer des concepts connus. Ce système est viable parce que l'analyse linguistique est relativement pauvre et parce que la représentation à obtenir est déjà partiellement connue à partir de la base de connaissances. Le système HAVANE de Bosc est également capable d'extraire du texte d'une petite annonce une représentation qui permettra ensuite de répondre à une requête émise par un utilisateur. Le système DOXIS de Membrado réalise une indexation conceptuelle sur des comptes rendus en langage médical. Le texte est résumé par un ensemble de concepts. On obtient ainsi le sens général du texte et l'on peut le comparer avec une question éventuelle. L'analyse des textes est essentiellement sémantique. Les concepts et leurs contextes d'interprétation sont connus et enregistrés dans un dictionnaire.

DOXIS résume le texte **"Astrocytome pariétal. Artériographie carotidienne (iode) de profil. Temps artériel. Processus expansif avasculaire. Encorbellement par artère cérébrale antérieure et les branches de la sylvienne..."** par l'ensemble des concepts: **tumeur,télencéphale, imagerie, cou(artère), image(technique employée), aspect, cerveau(artère), anatomie(détail), anormal.**

On peut également citer ici d'autres applications comme les travaux de S. Gauch, le système Alexis d'ERLI. Ce dernier système est destiné à la gestion de dictionnaires complexes. Il comporte un module d'analyse de la requête dont nous avons déjà parlé. Complété par des modules particuliers, il devient un système multi-experts pour l'aide à l'indexation et à l'interrogation. Il ramène alors la question de l'utilisateur aux notions présentes dans le thésaurus.

En d'autre terme: il passe de la notion de **"mobilier contemporain"** à celle de **"meuble moderne"**, de **"mobilier en bois massif"** à la rubrique **"meubles en bois"**. Dans un autre exemple, il retrouve la notion **"d'augmentation de salaire"**, que celle-ci soit présente sous la forme de **"croissance du SMIG"**, ou dans **"les récentes majorations des appointements des techniciens..."**, etc.

Dans ces systèmes, par l'utilisation de dictionnaires organisés comme des bases de connaissances, on repère un certain nombre de concepts. Ces concepts sont ensuite utilisés dans des procédures de comparaison *ne comportant généralement* **qu'une seule étape inférentielle.** Mais, dans une deuxième catégorie de système, on observe clairement l'**enchaînement de plusieures étapes inférentielles.**

Dans le système IOTA, les termes de la requête primitive sont progressivement transformés lors d'un processus de reformulation. Il s'agit d'abord de choisir les concepts à reformuler, puis de déterminer les relations sémantiques à utiliser, enfin de fixer le niveau de reformulation à effectuer. On substitue progressivement à un concept, un ou plusieurs autres concepts localisés en utilisant des relations sémantiques. On modifie successivement les liens entre les concepts. Un point important consiste alors à déterminer et à contôler l'éloignement *autorisé par rapport au concept initial.* Dans ce système, l'éloignement est une fonction du niveau de connaissance de l'utilisateur. On observe une réalisation du même type à travers le système EXPRIM de David et Créhange.

Dans le système IOTA, la question **"papier"** est transformée en "supports papier nécessaires" par deux étapes de reformulation. Dans le système EXPRIM, "(photos d') **enfants** qui font de la **course à pied**, pour la revue **Mode et Travaux"** devient: **"aspect vestimentaire et loisir des enfants"** en deux ou trois étapes.

Le système SPIRIT [Fluhr et Radasoa] notamment dans sa version multilingue [EMIR de Fluhr 91], utilise également un thesaurus comme base de connaissance d'un domaine spécifique. Il s'agit donc d'un aspect différent de la reformulation par les documents pertinents. De façon standard, le comparateur recherche les intersections entre les documents de la base et la question en manipulant des mots normalisés et en *utilisant un modèle statistique.* Dans ce cas, le comparateur du système exploite également à partir de chacun des termes de la question, les différentes relations possibles: synonyme, spécifique, générique, traduction, etc.

Les relances successives d'une recherche s'accompagnent donc d'une reformulation automatique de la requête initiale. Cette reformulation invoque un thesaurus - ou base de connaissance spécifique du domaine -. Mais elle s'appuie également sur des transformations linguistiques de type morphologiques ou syntagmatiques, sur l'utilisation de connaissances sémantiques contenues dans un dictionnaire général de la langue.

La nature exacte de ce processus d'inférence dépend du formalisme adopté pour représenter les connaissances. S'agissant de réseaux sémantiques ou de thésaurus, des opérations d'activation par proximité sont généralement utilisées. Les noeuds et les arcs représentant des concepts du domaine. Une fois les noeuds de départ activés, l'activation se propage vers d'autres noeuds en suivants les liens établis et en respectant certaines contraintes: contraintes de distance, contraintes de branchement pour un trop grand nombre d'arcs, contraintes valorisant certains chemins privilégiés en fonction de méta-connaissances, etc.

4. Conclusion: d'une architecture fondée sur des systèmes multi-experts vers une architecture néo-connexioniste?

On voit donc apparaître depuis quelques années des systèmes conviviaux et intelligents. Ils tirent parti de recherches cognitives et tendent à reproduire l'ensemble du comportement d'un expert **documentaliste**. Lors de la recherche d'informations, ils sont guidés par des connaissances stratégiques. Ils prennent en compte une modélisation des comportements de l'utilisateur. Ils disposent généralement d'interfaces en langage naturel qui s'appuient sur des **procédures d'analyse linguistique**. Ils incorporent, d'une façon ou d'une autre, des connaissances d'un **spécialiste du domaine**. Ils regroupent ainsi, dans un montage complexe, différents modules simulant l'intervention d'intermédiaires humains: documentaliste, linguiste et analyste (cogniticien).

Aux Etats Unis de bons représentants de cette nouvelle génération de systèmes sont CODER de Fox ou I3R de Croft. CODER met en oeuvre des experts, pour la construction des modèles de l'utilisateur et de la requête, pour l'analyse et l'indexation, pour le choix de différentes procédures de sélection et pour l'utilisation du thésaurus. Ces experts communiquent par un tableau noir. I3R dispose d'experts, pour construire un modèle de l'utilisateur, pour établir un modèle de la requête, pour le choix des procédures de sélection (modèle probabiliste, techniques de clusterisation, navigation), pour inférer à partir de la base de connaissances les concepts reliés à la requête initiale. Un contrôleur règle l'activation des experts en utilisant un plan et un agenda.

En France, on peut par exemple citer DIALECT [Bassano et Mekaouche]; SPIRIT [Fluhr et Radasoa] ou IOTA [Chiaramella et Defude]. L'interface DIALECT comprend deux experts stratèges contrôlant sept experts spécialistes. Le premier multi-experts stratège est responsable de l'analyse de la requête initiale et des textes. Le second est responsable de la reformulation progressive de la requête. Mais DIALECT2 dispose également des experts pour l'indexation, pour la sélection et pour la reformulation de SPIRIT. Dans une nouvelle version de SPIRIT, l'ensemble des connaissances sur la reformulation a été représenté de manière homogène par des règles de production. Ces règles prennent en compte aussi bien des connaissances linguistiques (familles de mots) que des connaissances sur le domaine (règles de type thesaurus). Le déclenchement de certains sous-ensembles de règles peut être réalisé au moyen de méta-règles en cours d'élaboration.

Dans IOTA, le système expert utilisé est un système à base de règles de production permettant l'appel de procédures externes en partie droite des règles. A travers sa base de données à court terme (tableau noir), le système expert gère la communication entre différentes composantes qui peuvent éventuellement être elles-mêmes qualifiées "d'experts": gestion de la base de textes, gestion d'un lexique de la langue et de règles d'analyse linguistique, gestion d'un thesaurus. Les procédures d'analyse linguistique constituent elles-même un montage multi-experts.

On remarque que les systèmes récents comportent un nombre de plus en plus conséquent d'experts. On remplace les réalisations traditionnelles, monolithiques et procédurales, par un ensemble de "petits" experts spécialisés. On reporte les difficultés de fonctionnement sur la communication et le contrôle du dialogue entre les experts. Il devient donc de plus en plus difficile de faire coopérer et de mettre en place l'ensemble de ces spécialistes. Or, un nouveau courant "néo-connexioniste, explorant la simulation de réseaux neuronaux par machine, inspire déjà quelques recherches en informatique documentaire. Une proposition intéressante repose sur la gestion de tous ces spécialistes par une méthode analogue à celle utilisée dans les architectures connexionnistes [Desrocques 90]. On suggère alors une sorte d'hybridation entre les systèmes multi-experts et les systèmes neuro-mimétiques. A l'architecture multi-experts, on emprunte des "experts" de taille très restreinte dont les connaissances sont fournies par des spécialistes (linguistes, documentalistes, etc.). Des architectures connexionnistes, on retient les possibilités d'apprentissage et de gestion efficace d'un grand nombre de traits. La réussite de ces réalisations construites en partie autour des modèles connexionnistes est probable: une convergence certaine relie cette approche aux méthodes statistiques traditionnelles et aux techniques de reformulation et d'inférence exposées dans cet état de l'art..

Bibliographie

Bancilhon F., Richard P. "Managing text and facts in a mixed database environment", in "New applications of databases", Gardarin G. and Gelenbe E. (Eds.), Academic Press, New York, 1984, pp. 87-107

Barthes C. et Rommens M., 1987, "Stratégie de recherche d'information en ligne: système expert EURISKO", congrès IDT, Strasbourg Mai 87

Barthes C., Glize P. et Carputat B., 1985, "Un système expert en recherche documentaire multibase et multiserveur", RIAO 85, Grenoble 1985

Bassano J-C. "DIALECT: an expert assistant for information retrieval (an intermediary system for information retieval)", AI conference (Canadian Artificial Conference), 1986, Montréal, Canada

Bassano J-C. "Systèmes experts et systèmes documentaires intelligents", 7 èmes journées internationales sur les systèmes experts et leurs applications, Avignon (France), 1987

Bassano J-C., 1985, "Un système convivial pour la recherche documentaire", RIAO 85, Grenoble 1985

Bates M. "Information search tactics", Journal of American Society for Information Science, 30(4), 1979, pp. 205-214

Belkin N.T and al. "Distributed expert-based systems; an interdisciplinary approach", Information Processing and Management, 23(5), pp.395-409, 1987

Bobrow D.G et Bates M. "IRUS: Information retrieval using a transportable natural language interface", ACM SIGIR (sixth international conference research and development in Information Retrieval), 1983

Bosc P, Courant M et Robin S., "a user interface based on a simple natural language", ACM Conference, Pisa, Italy, 1986

Brooks H.M et Belkins N.J., 1983, "Using discourse analysis for the design of information retrieval interaction mechanisms, Sixth annual international ACM SIGIR Conference, Washington 1986

Charniak E. "A parser with something for everyone", in "Parsing natural language" King M. (Ed.), Academic Press, 1983, pp. 118-149

Chiaramella Y., Defude B., Kerkouba D. and Bruandet M-F. "IOTA: a prototype of an information retrieval system", ACM SIGIR conference, Pisa (Italy), 1986

Créhange M. and al. "Semantic of user interface for image retrieval; possibility theory and learning techniques applied on two prototypes", RIAO 88, MIT, Cambridge MA, 21-24 March 1988

Croft W.B. and Roger Thompson "I3R: a new approach to the design of document retrieval systems", Journal of american society for information science, 1987

Croft W.B., "An expert assistant for a document retrieval system", RIAO 85, Grenoble 1985, pp.131-150

Dachelet R., "Etat de l'art de la recherche en informatique documentaire: la représentation des documents et l'accès à l'information", Rapport INRIA 1207, 32p., 1990

David J-M. et Créhange M., "l'activité inférentielle de reformulation des requêtes dans le système iconographique EXPRIM", RIAO 85, Grenoble 1985

Debili F., Radasoa P. and Fluhr C. "About reformulation in full-text Information Retrieval System", RIAO 88, MIT, Cambridge MA, 1988

Defude B., "Different levels of expertise for an expert system in Information Retrieval", 8 th annual international ACM SIGIR conference in Reaserch and development in Information Retrieval, Montréal, pp. 147-153, 1985

Delcroix J-C. "Semantic relationships and general framework for expert information retrieval systems (state of the art and guidelines for specifications), ESPRIT 901, report T2/1, July 1986

Desrocques G., Archambault D. and Bassano J-C., "An associative neural expert system for information retrieval", RIAO 91, Barcelone (Spain), 2-5 April 1991

Euzenat B., Normier B., Ogonowski A. and Zarri G-P. "SAPHIR+RESEDA: a new approach to intelligent database access", Proceeding of the 9 th IJCAI, Vol2, 1985, pp855-857

Fluhr C; and al., "Multilingual access to textual databases", RIAO 91, Barcelone, Avril 1991

Fouquéré C., "Systèmes d'analyse tolérante du langage naturel", Thèse de doctorat, Paris 13, 1988

Fox E.A and al. "Implementing a distributed expert-based information retrieval system", RIAO 88, MIT, Cambridge MA, 1988

Fox E.A. "Development of the CODER system: a testbed for Artificial Intelligence methods in Information Retrieval", Information Processing and Management, 23(4), 341-366, 1987

Gauch S. and Smith J.B. "Intelligent search of full-text databases", RIAO 88, MIT, Cambridge, MA, 1988

Gross M. "Lexicon grammar and syntactic analysis of french", proceeding Coling 84, 1984

Guida G. and Tasso C. "An expert intermediary system for interactive document retrieval", Automatica, 19(6), pp. 759-766, 1983

Harter S. and Anne Rogers Peters "Heuristics for online information retrieval: a typology and preliminary listing", Online Review, 9(5), 1985, pp. 407-424

Hendrix G.G. and al. "Developping a natural language interface to complex data", ACM Transaction on databases, 3(2), 1982

In-Search, "Adds BRS", professional version, Information Today, 2(1), January 1985

Kayser D., Karoubi M. et Nicaud L., 1985, "Modéliser le raisonnement sans le figer", Cognitiva 85, Paris 1985

Kayser D., 1982, "An experiment in knowledge representation", Proc. ECAI 82, Orsay 1982

Lallich-Boidin G., "Analyse syntaxique automatique du français, application à l'indexation automatique", thèse de doctorat, Grenoble, 1986

Lancel J-M., Simonin N., "Tex-Nat: a tool for indexing and information retrieval", RIAO 88, Cambridge MA, March 1988

Lebowitz M. "RESEARCHER: an experimental intelligent information system", procceding of the 9 th IJCAI, Vol. 2, 1985, pp. 858-862

MacLeod I.A. and Crawford R.G. "Document retrieval as database application", Information technology; Research and Development, Vol. 2, 1983, pp. 43-60

Marcus R.S. "Design questions in the development of expert system for retrieval assistance", Proceeding 49 th annual meeting American Soc. for Inf. Sci., Chicago, 1986

Marcus R.S "An automated assistant for information retrieval", Proceedings of the 44th ASIS annual meeting, Vol. 18, 1981, pp. 270-273

Marcus M.P. "A theory of syntactic recognition for natural language", MIT Press, 1980

Meadow C.T, and al. "A computer intermediary for interactive database searching", Journal of the ASIS, 33(4), 1882, pp. 325-332

Mekaouche A. et Bassano J-C., "Multi-experts system for documentary research", RIAO 91, Barcelone, Avril 1991

Mekaouche A. et Bassano J-C., "Analyseur linguistique multi-expert pour la recherche documentaire", Journées internationales Avignon 91, 1991

Membrado M., "Génération d'un système conceptuel écrit en langage de type semi-naturel en vue d'un traitement de données textuelles; Application au langage médical", thèse de doctorat, Paris 11, 1989

Miranda S. and Vintrou L. "Extension on a relational DBMS to handle texts in a micro-computer environment", Special workshop on new applications of DB, Churchill College, Cambridge, 1983

Moulinoux C., "Messidor: terminal d'interrogation de bases de données documentaires réparties", thèse de troisième cycle, Paris 11, 1984

Normier B. (ERLI) "French yellow pages, accessing heading via nomenclature in natural language", RIAO 88, MIT, Cambridge MA, 1988

Paice C.D., "Expert systems for information retrieval?", ASLIB Proceedings, 38(10), 1986

Pollitt A.S. "CANSEARCH: an expert systems approach to document retrieval", Information Processing and Management, 23(2), 1987, pp. 119-136

Radasoa H-P., "Méthodes d'amélioration de la pertinence dans un système de bases de données textuelles", thèse de doctorat, Paris 11, 1988

Rouault J. "Linguistic methods in information retrieval systems", Advances in intelligent retrieval: informatic 8, London:Aslib, 1985

Rouault J. "Linguistique automatique: applications documentaires", Berne, P. Lang, 1987

Sabah G. and Rady M. "A deterministic syntactic-semantic Parser", Proceeding of IJCAI 83, 1983, pp. 707-709

Sabah G., 1983, "Un système de questions-réponses sur les rubriques professionnelles de l'annuaire électronique", Publication 31 du GR 22, Paris 6, 1983

Salton G. and McGill M.J. "Introduction to information retrieval", McGraw-Hill, New York, 1983

Salton G. and al. "Extended boolean information retrieval", Communication ACM, 26(11), 1022-1036, 1983

Sager N., "Natural language information formatting: the automatic conversion of texts to a structured data base", in Yovits M.C. (Ed.), Advances in computer, 17, New York: Academic Press, 1978

Schank R.C, Kolodner J.L and DeJong G. "Conceptual information retrieval", in "Information retrieval research", Oddy R.N an al. (Eds.), Butterworths, London, 1981

Sellami M., "SMARD: un système multibase d'aide à la recherche documentaire", thèse de doctorat, Montpellier, 1988

Smith L.C. "Artificial intelligence in information retrieval systems", Information Processing and Management, 12, 1976, pp. 189-222

Smith L.C. "Artificial intelligence applications in Information systems", Annual review of information science and technology, American Society for Information Science, Vol. 15, 1980, pp. 67-105

Stout C and Marcinko T., "SCI-MAT: a menu driven universal on-line searcher and personal data management", Online, 7(5), September 1985

Tong R.M and al. "RUBRIC: an environment for full text information retrieval", Proceeding 8 th annual international ACM SIGIR conference in Reaserch and development in Information Retrieval, Montréal, 1985

Vickery A. and Brooks H.M. "PLEXUS:the expert system for referral", Information Processing and Management, 23(2), 1987, pp. 99-117

Walker G. and Janes J. "Expert systems as search intermediaries", Proceedings of the 47th annual meeting, Vol. 21, 1984, pp. 103-105

Wilks Y., 1979, "Making preference more active", Artificial Intelligence 11

Williams P.W. "The design of an expert system for access to information", proceedings of the 9th Int. Online Inform. meeting, 1985. Oxford and New Jersey: learned Information 1985, PP. 23-29

Zarri G-P., "Conceptual representation for knowledge bases and intelligent information retrieval systems", ACM, Grenoble, France, 1988

SEARCH STRATEGIES IN NATURAL LANGUAGE

Ir. B.H.A. Zijlstra
TDCK, Postbox 90701
2509 LS the Hague, NL

Summary

After a discussion of online searching problems some methods for making online searching easy for end-users are described : Intelligent gateways, ZOOM, HYPERLINE, CD-ROM and MENUS. Attention is given to parsing and natural language interfaces to databases and then natural language projects such as CITE, OKAPI, PLEXUS/TOME/MITI, DIANEGUIDE/NL, DGIS/STINET, SPIRIT/EMIR and Alpha DIDO are described. In the fourth chapter attention is given to Natural Language and Thesauri, such as the bilingual NATO Thesaurus. A bibliography has been added. In an annex an example shows that End-users can start online searching with Natural Language terms, using ZOOM and HYPERLINE commands.

1. THE PROBLEM

The problems with online searching (bibliographic) databases in their native mode on the commercial and governmental database vendor systems include :

What relevant databases exist
(which databases)
How do I access them
(which host) (SORM89)
How do I retrieve information from them (MISC87)
(which search terms)
(which search strategy)
What can I do with the retrieved information
(postprocessing) (COTT88)

1.1 WHICH DATABASES

Each hostcomputer has a guide which gives information about the databases that are available. Dialog has a Database Catalog, which mentions 7 databases about Education within the category Social Sciences and Humanities. But the Easynet Database Directory of Telesystems mentions 22 databases about Education, although PsychINFO and Social Scisearch are not mentioned. And Ted Brandhorst of the ERIC Processing and Reference Facility mentions even more databases about Education (BRAN90).
The I'M Guide of the CEC IMPACT Program lists 77 databases in the Category Education.

But when you have a question about a multidisciplinary subject such as "Interactive Videodisc for Education and Training", some of these Education databases will not have information about the subject and databases that are not mentioned may have valuable information. In that case online Database directories such as Dialog DIALINDEX and ESA QUESTINDEX may help.

```
DIALINDEX (SF engineering)

interactive(w)videodisc
INSPEC           322
NTIS              89
COMPENDEX PLUS    71
```

These database directories are hostspecific. When you want to know everything about a subject, you have to use all the database directories of the hosts to which you have access.
The I'M Guide database of ECHO, which contains information about some 1500 European databases isn't hostspecific, but is of limited use for a multidisciplinary question. The printed I'M Guide only gives information at the level of Dialog Bluesheets, which describe the format of the databases. A better tool might be the Online Manual of Blackwell Publishers, which contains a new and powerful keyword Thesaurus, designed to enable any user to find the most useful database to search in any given subject (COUS91).
But for a multidisciplinary question a Current Contents based Database Directory (CCDD) might be a solution.

1.2 WHICH HOST

In general you can get access to databases on hostcomputers if you have a PC, a modem, communication software and contracts with several hostcomputers. Unfortunately each hostcomputer has its own command language and a complicated logon procedure (Table I).

CCL

In 1979 the CEC developed a standardized common command language (CCL) as a tool to improve human utilization of computer-based information systems.

Table I

HOST	DIMDI	FIZ-TECHNIK, DATA-STAR	STN INTERNATIONAL	GENIOS	GEM	IW KÖLN, JURIS	DIALOG	ESA
Sprache	GRIPS/DIRS	DSO	MESSENGER	STAIRS/MIKE	STAIRS	GOLEM	DIALOG	ESA-QUEST
Datenruf-Nr. des Hosts (NUA)	45221040006 ODER 45221040104	456900 10552 (FIZ) 228464110115 (DATA-STAR)	45724740211.X 45724740199.X	1. 45400030296 2. GENIOS	45690040117	45221090171	31104150002000 oder 31104150004800	2222620021 234219201156
Datenbank-auswahl, -wechsel	BASE Datenbankname (BAS)	Datenbank-kurzname ..C Datenbank kurzname	FILE-Datenbankname (FIL)	..SEARCH Datenbank-kürzeb (..S)	Direkteingabe Datenbank-kurzname	Menü	BEGIN Datenbankar. (B)	BEGIN Datenbankar. oder name (B)
Datenbank-überblick	BASE	..C NEWS	HELPFILE NAMES	..INFO	..H BASES	-	B230 (Datenbank der Datenbanken)	?FILES
Eingabe Suchbegriff	FIND (F)	..SEARCH (..s)	SEARCH	Direkteingabe	..SEARCH (..SEA)	FINDE	SELECT (S)	SELECT (S)
Maskierung	$	$	?	S	?	&&	?	?
Suche eines Begriffes in bestimmten Feldern	FTI = Mikro-computer F Mikrocomputer (TI:AB)	Mikrocomputer. TI. Mikrocomputer TI.AB.	S Mikro-computer /TI S Mikrocomputer/ TI.AB	Mikro-computer TX.	..SEA Mikro-computer. Titel	-	S Mikro-computer/AB. TI S AU = ROSS.E?	S Microcomputer/ Ab. TI SAU = Bard?
Anzeige gefundender Dokumente	SHOW (S) (Eingabefolge: S Suchschritt-Nr. F = Felder; R = Dok.-Nrn.)*	..PRINT (..P) (Abfrage folgt)	DISPLAY (Eingabefolge: D Suchschritt-Nr.; Felder; Dok.-Nrn.)**)	..PRINT (..p) (Eingabefolge: ..P Dok-Nrn./ Paragraphen/ Suchschritt-Nr.)**)	..BROWSE (..B) (Eingabefolge: ..B Suchschritt-Nr. Doc=Dok. Nrn. Format)**)	ZI oder text **)	TYPE (T) (Eingabefolge: T Suchschritt-Nr./ Format-Nr./ Dok.-Nr.)	TYPE (T) (Eingabefolge: T Suchschritt-Nr./ Format-Nr./ Dok.-Nr.)
Zeigen der gesamten Profiltabelle	TAB	..DISPLAY all (..D)	DISPLAY HISTORY (D HIS)	..DISPLAY all (..D)	..DISPLAY (..D)	FRAGE	DISPLAY SETS (DS)	DISPLAY SETS (DS)
Suche im Thesaurus	DISPLAY (D)	ROOT	EXPAND (E)	-	-	F Suchbegriff**	EXPAND (E)	EXPAND (E)
Fortsetzung der Liste mit	MORE	ROOT*	E	-	-	ja/nein	PAGE	PAGE

Ubernahme v. Thesaurus-Begriffen in die Profil-tabelle	Eingabe der
	Tab.-Nr. z. B.
	1.01; 1.02; 1.05
		Eingabe der
		R.-Nr. z. B.
		R2-R6
		R2 R3
			Eingabe: SE-Nr.
			z. B.
			SE8-E15
			SE1 or E4 or E7
				-	-	1. Markieren des Suchbegriffes
						mit ‡‡
						Eingabe d. Nr.
						2. FRAGE
							Eingabe:
							SELECT E Nr.
							z. B. SELECT E7
							SELECT E1 or E8
								SELECT E Nr.
								z. B.
								SELECT E4
								SELECT E5, E7

Dialog beenden	STOP	..OFF	LOGOFF (LOG)	..OFF	..OFF	1. ENDE 2. LOGOFF	LOGOFF	LOGOFF
Benutzer-hilfen	*HELP* (?) ? Befehl INFO INFO	-	HELP (?) ? Befehl	..HELP (..H) ..H Befehl	..HELP (H) ..*H functions* ..H bases	? oder ??	EXPLAIN (?)	? ? Commands ? Databases
Datenbank-übergreifende Recherche	BASE Datenbankname; S = all Datenbankwechsel	..C CROS	-	..CROSS (..CR) ..CR Blockname ..CR alle	-	PUSSY Datenbankname	1. Über alle Datenbanken B 411	-
FREQUENZ ANALYSE								Z (ZOOM)

Table II

gateway on	Autodial Communicat	Command translation	Database selection	Search strategy	Natural language
Hostcomputer Mainframe	ESA-CODUS	Westlaw-Dialog	Questindex Dianeguide	CONIT	Dianeguide
Gateway computer Mini	Easynet II DGIS	Easynet II DGIS/CCL	Easynet II DGIS	QBES DGIS	QBES DGIS
PC of end-user	A-COM PSI Dialoglink	A-COM PSI TOOTSI	PSI	TOME MITI WilSearch	TOME MITI

Unfortunately the hostcomputers preferred to continue using their own command languages.

It is very useful to have an autodial modem which can store the logon procedure. Some hostcomputers have their own communication software, such as Dialink of Dialog and Mikrotel of ESA. Some hostcomputers have a gateway to another hostcomputer : When you have access to the ESA hostcomputer, you can also search databases on the Profile computer. But when you want to have access to databases on more than 3 hostcomputers, with different command languages, it might be useful to have an (intelligent) gateway or intelligent interface.

(Intelligent) gateways can help untrained end-users and information specialists to get access to bibliographic databases. Several possibilities exist, which depend on which computer the gateway is situated (hostcomputer, special gateway computer or PC of end-user) and the degree of intelligence or help (autodial telecommunication, command translation for different hostcomputers, database and host selection, terminology and search strategy help, search reformulation. (WILL86) (BOUM90) (EFTH90). (Table II)

The Easynet system of Telebase, which gives easy access to some 900 scientific, technical and business databases on 12 hostcomputers has been a commercial success since 1984. In this case you don't have problems with the command languages and the logon procedures of the hostcomputers. You do not need contracts with hostcomputers, you only get a bill of Easynet.
Intelligent Information (II) of Infotap which was developed later in Europe is giving easy access to business databases. Infotap is also developing TOOTSI, a tool-box for developing intelligent gateways, which will be available in 1992 (MAHO90).

Although there isn't much information about how Easynet processes a question, the Easynet system might be based on the ideas that were developed in 1975-1980 by Richard Marcus (CONIT) and Victor Hampel (TIS).

The ESURS system for automatic translation of command languages of the Dialog, Datastar, Fiz-Technik and Genios hostcomputers is based on lexical analysis, syntactic analysis (top-down parsing), semantic analysis and a generating proces (ZBOR91).

The major problem with intelligent interfaces is that, like the host menu-systems, they offer insufficient assistance with term selection and strategy construction. Simple searches may produce good results, but complex searches will probably be less satisfactorily resolved. Much will still depend upon the user's skill in selecting the right terms and using them correctly (LARG90).

1.3 TERMINOLOGY AND SEARCH STRATEGY

When you want to retrieve information about a certain subject from a database you have to use the Keywords from the Controlled Vocabulary or Systematic Thesaurus of the Database. That's why an Information Centre normally has Thesauri and Classifications of all databases that are used regularly : NASA, INSPEC, ERIC, PsychInfo, DTIC, TEST (for NATO-PCO database).

To retrieve information about a complex subject, the query has to be analysed and translated into keywords that have to be combined in a search strategy or search profile (SORM89) (BATE89). An overview of the various types of search strategies has been given by Harter (HART86).

BUILDING BLOCK STRATEGY

This is the most commonly used overall approach. A search profile is created through four steps :
- Identify major concepts or facets and their logical relationships
- Identify search strings that represent the the concepts (words, phrases, descriptors,identifiers,classification codes) and fields to be searched.
- Create a set for each concept or facet by combining the research strings of a concept using Boolean operator OR (union)

- Create a result set by combining the facet (concept) sets with Boolean AND (intersection) (NOT or OR rarely used)

An example of a search has been taken from a report about "Interactive Videodisc Instruction" (FLET90).

Only DTIC, ERIC and PsychInfo databases were searched, using all combinations of the following :

Block 1	AND	Block 2	AND	Block 3
Computer		Assisted		Education
OR		OR		OR
Videodisc		Aided		Learning
		Mediated		Training
		Managed		
		Based		

Additionally, the following terms were used by themselves :
Interactive Videodisc, Interactive Video, Interactive Courseware

CITATION INDEXING STRATEGY

This strategy is only applicable in the Science Citation Index and other databases based on citation indexes. The simplest strategy is to identify a "classic" highly relevant paper and to retrieve all the documents that have cited this document. Also the name of a cited author or the names of cocited authors can be used as a search criterion.

CONFERENCE CITATION INDEXING STRATEGY

When you can find a good conference about a certain subject, you have found key-authors of this subject. And then you can search all relevant databases using the names of these key-authors. Cited references to their articles can be found in the Science Citation Index.

1.3.1 **WHICH SEARCH TERMS**

The Easynet Database Directory of Telesystems shows that 22 databases have information about Education. When all available information concerning a certain subject is needed for a new research-project, several of these 22 databases should be searched. Dialog Onesearch or ESA Clustersearch or the Datastar Starsearch option can be used, which allows for multiple database searching with one strategy. Generally for Onesearch and Clustersearch long search strategies have to be used, because each database has its own Thesaurus, and relevant descriptors of all relevant databases should be included. A descriptor of a certain database will be a free term (Natural Language term) for another database (FIDE86). Alternatively, a strategy translation step between systems is needed.

ZOOM AND HYPERLINE

ESA-QUEST has an unique facility called ZOOM, which helps you to find useful descriptors. It is based on statistical characterization of terms associated with specified document sets. This is in general a Frequency Analysis of terms in various fields of the retrieved documents. ZOOM thus provides data for establishing statistical relations among terms, and between documents and terms. The information derived from ZOOM, in combination with general statistics of the database, can be used for Probabilistic Relevance Feedback and in other semi-automatic feedback modes (BELK90).
If the term the user is entering is not in the thesaurus, then a sample of the documents containing that term is examined. The controlled keywords (Descriptors or controlled terms) of the documents in the sample are ranked according to their frequency in the sample of documents using the ZOOM command (INGW84). Then you can reformulate the question.

ZOOM s interactive(w)videodisc

	Inspec		NTIS
	abstr	tit	title
number of hits	252	163	35
ESA ZOOM on	50	50	35
video and audiodisc	37	44	
videodisks			12
videorecording			6
interactive systems	31	39	4
interactive videodisc	21	15	
interactive video	7	5	21
interactive videodisc technology	5		

The ZOOM Command has been used for the HYPERLINE Searching aid, which contains a Browse Thesaurus option (Semantic Association Method) and a Navigation method. In the Browse Thesaurus option the top five controlled keywords that are also Thesaurus entries are shown to the user. Phrases (multiple terms) are accepted in input, but no real Natural Language Processing (NLP) is performed on them. This means that no verbs should be used and that (at least at the stage of the project) only a simple preprocessing is performed. In this preprocessing stage terms like "and" and "or" are treated as logical ANDs ar ORs and the term "in" is transformed in a logical "AND". The kind of phrases the user can input are then analogous to the one the user can search for via the CCL command "Find" (AGOS91).

```
Search term :
SGML

HYPERLINE BROWSE

            Related Thesaurus terms :

Electronic Publishing
Standards
Word Processing
Desktop Publishing
Electronic Data Interchange

Search term :
Electronic Data Interchange

HYPERLINE NAVIGATION :

 4994 : Data Handling
  993 : EFTS
 5974 : CAD/CAM
  763 : Integrated Software
39674 : Standards
  109 : EDIF
```

ZOOM and HYPERLINE have much to do with the quorum function that was proposed by Cleverdon (CLEV84) and which was implemented on a trial basis by ESA-IRS in 1985 as Questquorum (HART90).

BRS TERM AND SUPERTHESAURUS

But if all the cooperating systems used a common set of search keys (descriptors) from a single generic (faceted) Thesaurus the translation step between systems could be avoided (VICK75).

The uncoordinated growth of databases is an impediment to online searching. Standardization and coordination are required if users are to be able to fully exploit the capabilities of on-line systems (FIDE88). Standardization and improvement in Subject Access to recorded information worldwide must be accomplished before Expert systems can be truely useful for information retrieval (ALBE90)

The BRS hostcomputer has a TERM database, which contains merged Thesauri from several fields (Education, Psychology, andMedicine, amongothers), as well as Natural Language terms suggested by practicing searchers. Even suggested Boolean combinations are included. When a searcher inputs a term or phrase into the BRS TERM database, an entry is printed out listing possible alternative terms that may be up to several dozen lines long.
If this database were expanded and enriched as a front-end database, or a Superthesaurus, it could contain an enormous variety of useful entry terms, with all sorts of guidance to decide on the best terms for a given search (BATE89).

1.3.2 WHICH SEARCH STRATEGY

BOOLEAN OPEARATORS

During online searching Boolean operators AND, OR, NOT are used to combine the different aspects/concepts of the question (postcoordination of descriptors). There is much comment on the difficulty people have with Boolean operators in database querying (THOM89), (ESSE91). Borgman found that 25 % of subjects learning an SQL-like query language could not pass benchmark tests for system proficiency, although these tests were representative of the searches that were supported. The problem seemed to lie in the use of Boolean logic : more than one quarter of the subjects could not complete simple search tasks involving the use of one index and at most one Boolean operator (BORG86).

One aspect of this problem might be the difference between natural English usage of the words "and" and "or" and their use as Boolean operators in retrieval. The English word "or" is most frequently used to indicate union, but "and" is often used ambiguously to indicate both union and intersection. The other aspect might be that the subjects do not understand the use of Boolean operators in subset specification.

It seems that Boolean operators are the biggest problem in End-User online searching. Ulla de Stricker even doubts if a menu interface can be designed which will mediate between the "online innocent" End-user and one or more online services (STRI88). But a menu interface is used in Urbana to construct a Boolean search strategy (MISC89).

Frants and Shapiro describe an algorithm that automatically constructs a query. The algorithm uses a set of documents that the user found pertinent to his information need. The descriptors which were assigned to these documents are analyzed, and the most important descriptors are identified. Then the Boolean query formulation is drawn up, consisting of subrequests, of which the calculated values all exceed a certain bound value. This bound value determines the quality of the query formulation (FRAN91).

ACCRUE (RANKING) OPERATOR AND QUERY-BY-EXAMPLE

Instead of the "difficult" Boolean operators a more user-friendly ACCRUE operator is being used in Topic, a full-text retrieval software, which is based on the RUBRIC concept-tree, a faceted cluster of 50 terms (TONG85). The ACCRUE operator works as follows. Suppose you want to search with 4 keywords A,B,C,D in a database of abstracts which contain 4, 3, 2, 1 or 0 of the 4 keywords.
With the ACCRUE operator titles related to abstracts that contain 4 keywords come on top of a list, then abstracts that contain 3 keywords and so on.
The titles having the same number of keywords in the abstract are ranked according an individual weight factor for each keyword. The End-User does not have to think about Boolean operators, the ACCRUE operator will get the most relevant titles on top of the list.

The ACCRUE operator can be used for Query-by-Example. In this case you start a search with an abstrct of a known very relevant article. You just click some 5 relevant words in the abstract and then the ACCRUE operator will find similar articles, with the titles of the most relevant articles on the top of the list.

RELEVANCE FEEDBACK AND WAIS

Instead of Boolean operators a new Relevance Feedback concept has been developed as retrieval method for terabyte databases for the Connection Machine supercomputer of Thinking Machines (KAHL86). The Relevance Feedback concept has been used as DOWQUEST for the Dow Jones News database on a CM2 machine. First a seed search is made by entering all relevant words, then the say 10 most relevant articles of 100 are marked and then a search is made for "similar" articles. The Relevance Feedback concept has also been used for Wide Area Information Servers (WAIS). WAIS is a standard information exchange protocol that offers unlimited connectivity and retrieval functionality (KAHL91).

PROXIMITY OPERATORS AND RELATIONAL KEYWORDS

Powerful proximity operators such as "adjacent" ADJ or (Ow) can be used for Natural Language searching for non-descriptors in Full-text retrieval systems : Intelligent ADJ Information ADJ Retrieval.
Because words like with, of, from etc. are on a stop list, you can not use proximity operators for those words. In this case you can use relational keywords (MCGI89).

Software-development-wi-Computer-graphics
Specifications-of-Software
Aircraft-vs-Air-Defence

2. CD-ROM AND MENUS

In 1985 Philips and Sony published the "Yellow Book", a loose standard for using the audio CD for digital data. This allows to store some 600 Mb of text information on a compact disc. The disc thus became a Read-Only Memory which can hold some 200.000 pages of text.
In 1985 Silver Platter, which was founded by Bela Hatvany, presented the first commercial CD-ROM offline database product. Since 1985 CD-ROM products have become a success, because they generally are easier to use than online databases and because you have "unlimited" access time, once you have a CD-ROM of a certain database. In 1991 some 2000 commercial CD-ROM products exist.

The NTIS Ondisc of CD-ROM Market-leader Dialog is a most useful CD-ROM product in Aerospace R&D, because it contains much information about Defense Research reports (AD-A nrs) and NASA research reports (N nrs).
It has an Easy Menu mode which is really user-friendly and a Dialog Command Mode. In the Easy Menu mode first you enter terms (OR) and then you limit (AND).

A search strategy can be saved and executed online, even Easy Menu strategies. This permits the End-User to search online. Several Easy Menu Commands can be used in the Command Mode.
In the display mode of the Dialog Ondisc sorting on Word Frequency allows for Ranking of the "Best" on top, even when the search strategy is clumsy. This is an useful option for untrained End-Users, who can also use the Dialog Ondisc for training Online searching. Dialog produced their first CD-ROM in 1987, the ERIC Ondisc, and now offers 35 CD-ROM products.

But some other CD-ROM products also have nice things.

- NATO-PCO CD-ROM has an user-friendly Form Filling mode
- Silver Platter products have a F2 Functionkey which you can use to transfer words from the abstract to the Search function.
- WilsOndisc of Applied Sciences and Technology has a rather good Thesaurus

Unfortunately each CD-ROM vendor uses his own software. In this case you can decide to standardize on 1 vendor who offers several CD-ROM databases. In a big information centre with several Information Specialists you can also decide that Information Specialist 1 makes use of Dialog CD-ROM's and nr 2 makes use of Silver Platter CD-ROM's. But in the case of a small Information Centre with 1 Information Specialist for all kinds of databases you would still need an intelligent interface.

In september 1991 Silver Platter announced to split their SPIRS software in a Database Search Engine (server) and an Interface (client), according to the recommandations of the NISO Standards Committee (This would maybe allow users to search Silver Platter CD-ROM products with the Dialog retrieval software, but maybe less information can be stored). Silver Platter is also working on the Electronic Reference Library (ERL) for Macintosh and Sun computers.
Standards are being developed for :
- Information Retrieval Protocol (Z39.50)
- Structured Full-Text Query Language (Aerospace Industry)
- CD-ROM Read-only Data Exchange

In september 1991 Dialog announced the CD-ROM version of the Bluesheets which contains information about the format of the online databases. The Bluesheets OnDisc can also be used as a simple database directory.

But at this moment CD-ROM's do not have a good frequency analysis such as the ZOOM command of ESA.

NATURAL LANGUAGE SEARCHING ON CD-ROM

Because titles of articles contain Natural Language and are searchable on a CD-ROM, you can start a search with a part of the title of a good article.

GLOBAL CD-ROM : EDUCATION

In case intelligent gateways will give easy access to databases in the whole wide world, a global information system will appear (KRAN89) (RICH89).
But instead of intelligent gateways to all the databases about Education,

it might be easier to combine all the Education information from all relevant databases on 1 single global Education CD-ROM that contains max 5 years, yearly updated.
(The ERIC CD-ROM now is max 10 years)

MENUS :

The creation of new business information services by Compuserve and Dow Jones has been an important factor in the world-wide market for business information. Their strategy was to target the End-user by offering easy-to-use software and an emphasis on business news. Dialog attempted to get into the game by offering the Business Connection, an easy-to-use interface that stands between the user and Dialog's more complex command driven software (BREM91).

DMC

The Dialog Medical Connection (DMC) gives menu-driven access to 30 databases such as Biosis, Medline, Scisearch and Excerpta Medica. When 91 biologists and chemists of the Biology Directorate of Glaxo Group Research Ltd got access to the DMC, the number of searches increased 200 %. The average End-User carries out 4 searches per month, spending 1 hour and some $ 60 (BOYD90). Of 4 alternatives (Searcher, IT, Easynet with its high pricing structure, DMC) DMC matches most of the requirements of the End-Users and is tailored to the needs cf biomedical searchers. DMC is an acceptable and effective system for End-User searching for biologists and chemists within Glaxo Group (BYSO90).

The success of DMC and CD-ROM Menus have been a driving force for Host computers to make their databases also available through Menus.
Several Hostcomputers now have come with Menus for Easy online access, but users might prefer a Common Menu Mode (CMM) for both CD-ROM Menus and Online Menus.

3. EXPERT SYSTEMS AND NATURAL LANGUAGE

Expert Systems have great potential for Information Retrieval. Expert Systems contain Knowledge Bases (Thesauri etc.) and Rules. In case Expert Systems will be used to help End-Users, the rules which operate on the Knowledge Base should be analogous to the decision rules or work patterns of the Information Specialist (SHOV85). Expert systems can also help intermediaries. But end-users and intermediary interface needs are very different (COTT87). Or, where should the person stop and the Information Specialist search interface start (BATE90).

A truly intelligent computer interface would be one, then, which would perform all of the functions that an intelligent, successfull human intermediary does, which are necessary and sufficient for successful information retrieval system performance.
Furthermore, it would be necessary for these functions to be performed in an appropriate interactive dialogue, perhaps also based on that which is performed in human-human information interaction (BELK86).

Because Du Pont has discovered that expert systems can do 80 % or more of the decision work of experts, it may be similarly expected that expert systems could locate 80 % or more of the information required by an interrogator (FEIG88).
As much as 80 % patron information needs will likely be satisfied by the intelligent interface. The remaining 20 %, the difficult, unusual, hard-to-pin-down questions, will be referred to human intermediaries (KRAN89).

USER MODEL

Before you start an online search, you have to decide how much money will be spent on the question of a requester. You have to know how important the question is.
Will the requester be happy when he can select 5 relevant articles from a list of 10 titles or does he need to know everything from the whole wide world concerning the subject of a new R&D contract.
3 types of online users are : Novice End-users (NE), Subject experts (SE) and Online Experts (OE). Other types are described in literature about "User Behaviour" (BELK85) or "User Modeling".

Table III

ZOOM on AUTHORS in ESA.

(IR) AND Intelligent

INSPEC

384 hits :

- 7 Jones, K.P. (Ed)
- 6 Belkin, N.J.
- 5 van Brakel, P.A.
- 4 Benenfeld, A.R.
- 4 Brajnik, G.
- 4 Brooks, H.M.
- 4 Burton, H.D.
- 4 Guida, G.
- 3 Croft, W.B.
- 3 Vickery, A.
- 2 Chiaramella, Y
- 2 Clemencin, G.
- 2 Fox, E.A.
- 2 Jacobs, P.S.
- 2 Marchetti, P.G.
- 2 Pejtersen, A.M.
- 2 Pollit, A.S.
- 2 Sormunen, E.
- 2 Thompson, R.H.

Generally, it is allways useful to make an expand on the name of the author

(IR) AND Intelligent

Pascal 205
Pascal 204
NTIS
INSPEC

512 hits :

- 7 Belkin, N.J.
- 7 Jones, K.P. Ed
- 7 V.E. Hampel
- 6 Zarri, G.P.
- 5 Brajnik, G.
- 5 Brooks, H.M.
- 5 Croft, W.B.
- 5 Guida, G.
- 5 Rau, L.F.
- 5 van Brakel, P.A.
- 4 Benefeld, A.R. Ed
- 4 Burton, H.D.
- 4 Gladys, A. Cotter
- 4 H.D. Burton
- 4 Williams, M.E.
- 3 Chiamarella, Y.
- 3 Fox, E.A.
- 3 Gauch, S.
- 3 Hampel, V.E.
- 3 Jacobs, P.S.
- 3 Richard W. Hartt
- 3 Thompson, R.
- 3 van Rijsbergen, C.J. Ed
- 3 Vickery, A.
- 2 Clemencin, G.
- 2 Marchetti, P.G.
- 2 Pejtersen, A.M.
- 2 Pollit, A.S.
- 2 Roger, C. Schank
- 2 Sormunen, E.

Table IV

When we want to find key authors, we should look for the proceedings of good conferences about Intelligent Retrieval. In this case we might find :

		Authors			
1985	Advances in intelligent retrieval INFORMATICS 8. Proceedings of a conference sponsored by Aslib + Aslib Informatics Group	BROO	BELK	MITE	WALK
1986	Second Conference on Computer Interfaces and Intermediaries for Information Retrieval Boston, MA, May 28-31 1986 NTIS Nr AD-A174 000	BATE CROF MISC	BELK FOX SALT	BURT HAMP SMIT	COTT MARC TOLI
1987	ACM/SIGIR 10th International Conference on Research & Development in Information Retrieval New Orleans, LA, june 1987	BRUA	CROF	TONG	
1988	RIAO 88 User-Oriented Content-Based Text and Image Handling Recherche d'Information Assiste par Ordinateur MIT, Cambridge, MA, March 21-24 1988	BIXB POLL	FOUC RAU	LELU TONG	KUHN
1988	Online Information 88 12th International Online Information Meeting London, 6-8 December 1988. Proceedings, I+II Oxford : Learned Information	BERG	COTT	KUHN	
1989	ACM/SIGIR 12th International Conference on Research & Development in Information Retrieval Cambridge, Massachusetts USA, June 25-28 1989	CROF KWOK SALT	FUHR PARK STAN	HAAS PEJT	INGW RIJS
1989	12th International Conference on Computational Linguistics Budapest, Hungary, 22-27 August 1989	GEHR	HAUG		
1990	AGARD-TIP Bridging the Communication Gap AGARD-TIP Specialist Meeting Trondheim, Norway, 5-6 Setember 1990 AGARD Nr CP 487	LAWR	MAEG	ROLL	VICK
1990	ACM/SIGIR 13th International Conference on Research & Development in Information Retrieval Brussels, Belgium, 5-7 september 1990	BELK MARC	CROF ROUS	EGGH TONG	FOX
1990	Intelligente Schnittstellen zu Informationssystemen Darmstadt, 1-2 November 1990	AMME GREE	BÖCK KNOR	BOYL PEJT	CARB VICK
1991	RIAO 91 Intelligent Text and Image Handling Intelligent Multimedia Systems Barcelona, Spain, 2-5 April 1991	AGOS LIDD	BASS MARC	CROF RIJS	EVAN TOKU

Importance	$	1.000	10.000	100.000
Search budget	$	10	100	1.000
Document deliv	$	10	100	1.000
Searcher		NE	SE	OE
training hours		1	10	100
needed help				
which database		+	+	+
which terms		+	+	
which strategy		+		

This might mean that an intelligent *interface for 80 % of the questions* should mainly help Novice End-users (NE) and Subject Experts (SE).

In the case of the Novice End-User, the interface should be able to accept questions in Natural Language such as "Do you have something about XYZ", remove "do you have something about", split XYZ in components according the Building Block strategy etc.

In the case of an important question concerning a literature search for a new research contract with a value of over 100.000 $, where all available information is needed, a trained Online Information Expert should search. He will use the Building Block Strategy, cluster search etc. He will make a ZOOM on Journal names to find the names of the Core Journals, to be able to read the latest issues that are not yet searchable in databases. He will also make a ZOOM on authors, to find the names of Core Authors. (Table III). These names can be used for searching the Science Citation Index. The Online Expert will also search for information about Conferences, Symposia (Table IV).

Although intelligent interfaces can help Novice End-users and Subject Experts, I do not expect that intelligent interfaces will be able to make and interpret ZOOM *searches, searches* in the Science Citation Index, searches for Conferences and Symposia etc.

NATURAL LANGUAGE : I SAW HER DUCK

Natural language sometimes is not very easy to understand, because some of the words in a sentence may have different meanings. Saw can mean cutting with a saw or it can be the past time of the verb see. Duck can be a bird or a verb meaning bending. If the female person should live in the neighbourhood of an airport she might have made a bow because of the noise of a plane.

When you want to translate a sentence or answer a question, you must know what the sentence means or what the requester means. If you don't understand the sentence or the question, you have to ask a second question to check the context. In case the word "and" is used in the question "Do you have something about aeroplanes and meterology, you have to know whether the meaning is :

- union (aeroplanes OR meteorology) or
- *intersection*

(aeroplanes AND meteorology). Major problems in understanding natural language are : ambiguity, imprecision, incompleteness, inaccuracy (SCHA84)

PARSING

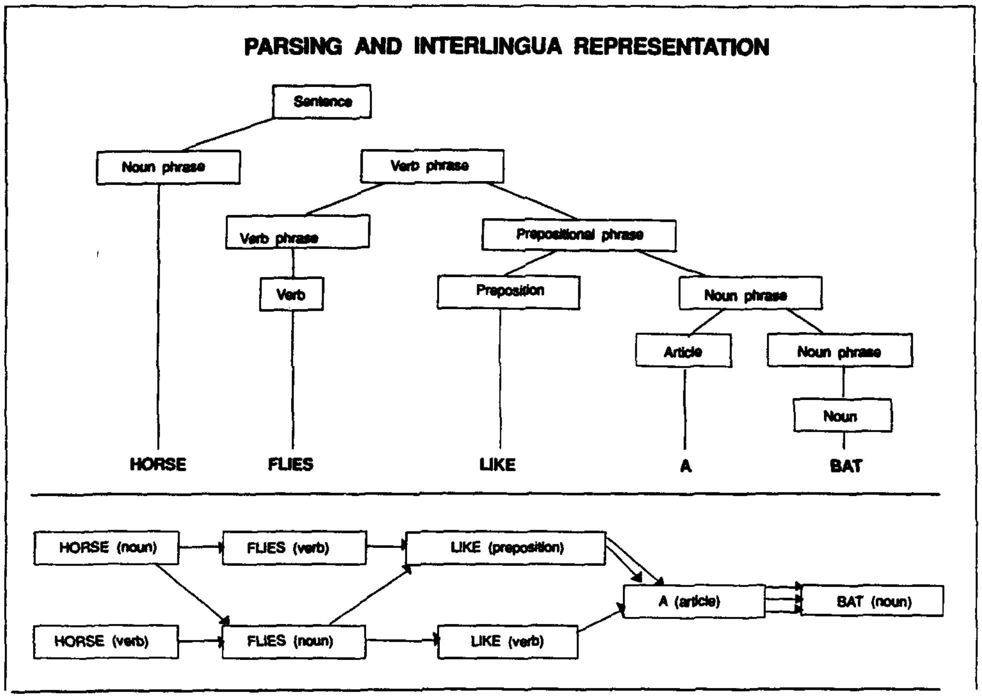
Parsing is a process that is used during automatic translation (ALBE90) and automatic indexing (EVAN91). During the parsing proces verbs, nouns and their logical order are recognised. This parsing process is also employed to understand the meaning of a question that will be processed in online database searching. It then has similarities with the process of splitting up a question into concepts or facets, which is employed in the online Building Block Method. Several types of parsers exist :

- Context-free parsers (late 50s and early 60s), which attempts to decompose a sentence by succesively applying a series of derivations such as :

1. sentence (S) = noun phrase (NP) + verb phrase (VP)
2. noun phrase = article (T) + noun (N)
3. verb phrase = verb (V) + *noun phrase*

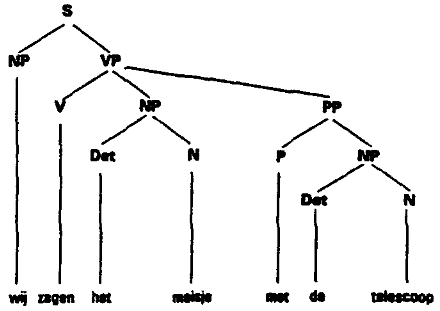
Bottom-up parsing starts at the level of the individual words. Bottom-up analysis uses knowledge about language and can produce accurate, if only partial results, in arbitrary texts and texts that contain unexpected information (JACO90).

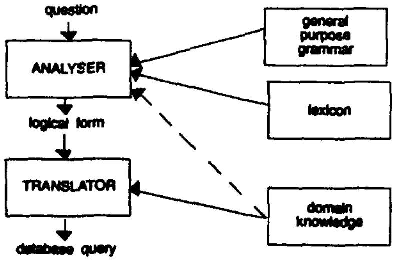
Top-down parsing starts at the top of the tree. Top-down analysis is much more tolerant of unknown words and grammatical lapses, but is often fooled or



Possible syntactic readings (parentheses added to clarify)

● (The HORSE (RUNS SO SWIFTLY, HE) flies like a bat.
→ ● HORSE FLIES (insects) LIKE (to feed on) A BAT.
● HORSE (around, you)FKIES LIKE A BAT.





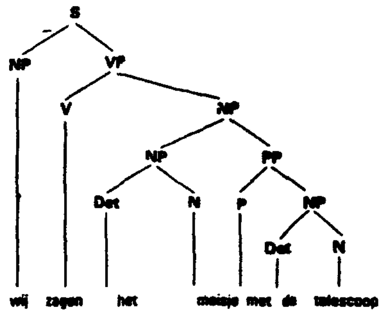


Figure 1

misses information in unusual situations.

- Transformational parsers (late 60s) generate the deep structure of a sentence, representing its syntactic and semantic interpretation.

- Augmented transition network (ATN) parsers (70s-) have the same power as a transformational grammar but are more straightforward (KORY90)

Copestake and Sparck Jones propose an analyser which carries out syntactic and semantic processing, based on a general purpose grammar and a domain-dependent lexicon, and a translator which is responsable for producing the database query. They use the term analyser rather than parser, because parsing is frequently taken to refer to syntactic processing alone (SPAR90).

Simple parsing yields a a single parse tree. even if the sentence is ambiguous and can be parsed several ways. At this stage the parsing is purely syntactic. Sophisticated parsing yields a parse forest composed of all parses that the grammar allows. Semantic analysis of the parse forest will yield a most likely interpretation, which becomes the interlingua interpretation.
An interlingua representation details the syntax of a sentence and includes enough semantics to increase the likelyhood of creating an accurate synthesis. Elementsof therepresentation are actually coded as numbers, that are indexes to multilingual dictionary entries and phrase structure templates.
The Distributed Language Translation system being developed by BSO Language Translation in the Netherlands uses Esperanto as its Interlingua (BENT91).

NATURAL LANGUAGE ACCESS TO DATABASES

Let's consider an untrained user in a library who wants information about Austria's clothing industry. In this case a search strategy for Dialog might be :

ss (clothing or garment or apparel) and
(austria or austrian)

Because the untrained user is unaware of the need to parse a query into component parts he enters the question "clothing industry in austria". In case the hostcomputer can recognize "in", an article "Austria sees clothing industry boom" would be missed because there is no word "in" in the title. In case "in" can be recognized and discarded as trivial word, the words "austria", "clothing" and "industry" can be combined by an automatic Boolean AND operator. Neither is satisfactory since the results of the search will not be consistent with the intention of the user (STRI88).

In this case a Natural Language Interface (NLI) or automatic Natural Language Processing (NLP) should translate the question in the above search strategy. In general NLI or NLP make the translation of the "free-form" natural language expression of information needs and queries by users to relevant system terms. NLP software may also construct the system query and trigger the query operation. NLP often exploits existing system indexes, thesauri and front-end dictionaries or semantic networks when performing one or more of these functions.

NLP linguistic analysis techniques have been used successfully in Online Public Access Catalogs (OPAC's) to assist with morphological (variant but equivalent word forms) and syntactical (variant but equivalent phrases) query-document matching problems.
For example, some routines compensate for word spelling or suffix variations; in this case the stemming algorithm of Porter may be used (PORT80). These algorithms were designed to conflate terms that are morphologically similar. The assumption is that they will be semantically close. Other match direct and inverted forms of subject headings. In some systems these approaches are combined to improve retrieval effectiveness (HILD89).

3.1 NATURAL LANGUAGE INTERFACES TO STRUCTURED DATABASES

There is a considerable body of literature on Natural Language (NL) interfaces to structured (non-text mainly) databases (FINI86) (NOAC90). There have been a number of Conference Panels on Natural Language Interfaces (NLI), such as the 20th Annual Meeting of the ACL, the 10th International Conference on Computational Linguistics/22nd Annual Meeting of the ACL and the 25th Annual Meeting of the ACL (WEIS87).
LADDER was developed by Hendrix in 1978, TQA by Damerau of IBM in 1980, IRUS by Bates in 1986, TEAM by Grosz in 1987 and JANUS in 1989. The TQA system is a NL front-end todatabases, whichtranslates NL queries into SQL expressions, which are then evaluated against the database. The Structured Query Language (SQL) is a database querying language codified as a standard by the American National Standards Institute. SQL is based on the Relational modeling of data in database management systems (DBMS). These new RDBMS systems (Oracle, Sybase, Ingres, Informix etc) are oriented to distributed application on local area networks with client/server architecture (CSA). Because querying directly with SQL is too complicated for non-technical people, an user-friendly interface is needed (LEIG89). EDA/SQL gives easy universal database access to DB2, Oracle, dBase, even MUMPS (RICC91).

Commercial NLI systems which have been sold more generally include INTELLECT (derived from Harris's ROBOT), which runs on IBM mainframes and is used by the Library of Congress to access its structuredpersonnelfiles (WARN88), Q&A (developed by Hendrix following the experience of LADDER) and Natural Language of Natural Language Inc., Berkeley (SPAR90).

Natural Language of Natural Language Inc can generate charts and graphs and employs syntactic and semantic analysis of a query with a Knowledge Base of over 100.000 English concepts and words.

SAPHIR from GSI ERLI is comparable with INTELLECT. GURU is a product that supports the development of expert system interfaces to a (relational) database management system. GURU also has capabilities for storing and manipulating rules of a sort that make possible the understanding of simple Natural Language queries (LEIG89).

INFOSTATION

VTLS Inc of Blacksburg has developed new generation software on a NeXT computer, the VTLS InfoStation, a multimedia access system for library automation. The multimedia software of Intermedia allows for Hypermedia links across documents containing text and graphics. An unique feature of the Infostation is its ability to understand Natural Language queries. An Expert System interprets user questions, sends requests to remote databases and even learns new vocabulary (LEE90).

SCISOR

SCISOR is a prototype Natural Language text processing system which was developed by Jacobs since 1986 for analysing stories about corporate mergers and acquisition from the Dow Jones database. The design of SCISOR combines lexical analysis and Natural Language analysis, knowledge representation and information retrieval techniques. The Natural Language part consists of a bottom-up parser TRUMP combined with a top-down parser TRUMPET.

A topic analyzer looks for some 150 prespecified keywords such as buy, merger etc. SCISOR uses a general lexicon of about 10.000 word roots, with links to a core concept hierarchy of 1.000 general conceptual categories.

SCISOR has been ported to the domain of military messages as a part of the MUCK-II project (JACO90).

NLH/E

Because menus and hypertext systems are unsatisfactory in help situations, Walter Tichy has built NLH/E, a Natural Language Help system, which answers questions that are entered as typed, Natural Language sentences. NLH/E is built with a novel caseframe parser that operates with a thesaurus, case inheritance, and noun/verb phrase unification (TICH89).

3.2 NATURAL LANGUAGE INTERFACES TO MULTIDISCIPLINARY DATABASES

When it comes to Natural Language Interfaces to multidisciplinary bibliographic databases and ill-structured full-text databases, generally other systems are mentioned in the literature, such as the CITE, OKAPI, ALEXIS/DIANEGUIDE/I'M GUIDE, PLEXUS/TOME/MITI, DGIS/STINET, SPIRIT/EMIR/DIALECT and Alpha DIDO projects.

I will first give attention to Online Public Access Catalogues (OPACs) in libraries, because OPACs are used by end-users that form a large population with extremely varied needs and backgrounds (MITE89).

The initial OPAC was derived from circulation or cataloguing systems and gives precoordinatedphrase access to separate indexes (subject headings, title, authors), which is appropriate for known item search. The second-generation OPAC systems are more similar to the commercial Information Retrieval systems; they provide keyword or free-text access, corresponding to post-coordinate IR principles. Their interfaces are generally command-driven, making use of Boolean logic which is well suited for subject searching.

The third-generation OPAC's combine the features of the first 'wo OPAC's by providing both phrase (known item) searching and keyword searching (MITE89).

3.2.1 CITE

CITE, the third-generation OPAC at the National Library of Medicine (NLM), supports Natural Language queries and performs intelligent stemming (truncation) on the user's search terms. Stemmed query words are looked up in both free text indexes and the MeSH (Medical Subject Headings) Thesaurus. Search words found in free text are then automatically linked to associated MeSH descriptors.

The CITE retrieval methods include term weighting, combinatoric searching, closest-match search strategy, relevance feedback (dynamic user feedback), query expansion, and ranked document output. Documents estimated to be most relevant are output at the top of the list.

To begin a search, CITE invites the user to "Type your search question". Significant words in the query (those not stoplisted) are passed through a customized stemming algorithm which conflates variant word forms, but also removes endings very common in medical vocabulary (e.g. -itis, -ectomy).

Even without employing semantic aids such as synonym tables, or a front-end dictionary or semantic network which would map potential search terms with related MeSH descriptors, CITE achieves a considerable degree of semantic query expansion using its automatic and semi-automatic processes.

3.2.2 OKAPI

OKAPI is a prototype third-generation OPAC developed by a research team (Mitev and Walker) at the Polytechnic of Central London. OKAPI-84, supported Natural Language subject searching, with weighted term, combinatoric retrieval. OKAPI-84 also employed search decision tree-based rules to automatically change search strategy when one attempt failed to produce retrieved items. Although the 1984 version had flexible retrieval routines built in, matching on entry words had to be nearly exact (known-item search).

OKAPI-86, a newer version, tested several linguistic computing techniques aimed at improved term matching : automatic "WEAK" and "STRONG" stemming (truncation) of search terms, automatic cross-referencing and semi-automatic spelling correction. The research team installed Porter's stemming algorithms in OKAPI-86. The "WEAK" stemming routine normalized singular and plural noun forms, and removed possessive endings, -ed's and -ing's. Spelling variations (e.g. UK and US forms) were incorporated into the weak stemming procedure. A stronger routine was tested but later rejected.

A measure of synonym control was attempted, specifically, automatic cross referencing, using a look-up table which had 3 kinds of entries : stop words, synonym pairs that would adversely affected by conflation (e.g. child, and children), abbrevations and their full expressions, words with alternative spellings (jail, goal) and equivalent word pairs suggested by query terms found in logged searches (Great Britain, UK).

OKAPI-86 OPAC corrected about half of the spelling/miskeying errors with favorable results. The cross referencing always helped when it was called into play. About 25 % of the searches studied contained a word or phrase that matched one in the cross references list. Very clearly, OKAPI-86 has shown that subject retrieval performance can be improved through the use of automatic linguistic term matching aids (HILD89).
Several functions of OKAPI have been integrated in the LIBERTAS system of Swalcap Library Systems (SLS).
A comparison study of LIBERTAS and OKAPI was published in 1989 (HILD89). LIBERTAS is the only "fourth-generation" OPAC commercially available that gives intelligent assistance and which ranks the result of a search.

3.2.3. PLEXUS/TOME/MITI

PLEXUS is essentially a Natural Language system, centered around the use of the facet classification scheme BSO, or Broad Subject Ordering (ALBE90).
PLEXUS was designed since 1983 by Alina Vickery and Helen Brooks as a prototype expert system for referral under a British Library R & D Department grant at the Central Information Service of the University of London. PLEXUS was designed to help librarians to deal with referral queries about Gardening. It has a dictionary of terms, a (knowledge) database of 500 records of referral resources (general gardening reference works, societies and experts) and a hierarchical classification of the subject domain "Gardening", by facets using the Broad System of Ordening (BSO).
Input to it is free form and can be Natural Language : a list of terms, a phrase, or a sentence. The user's input is passed through a parser to extract the significant words which are then used to construct a Boolean query for the database. (HAWK88). The search employs knowledge of search strategies and tactics to broaden and narrow the search.
The search strategy is modified automatically if the user wants more or less information. PLEXUS also constructs an user-model : by a series of questions it assesses the level of experience of the user (MITE89).
To handle input in Natural Language a stopword list and stemming rules were needed. The stopword list contains about 1400 terms : all articles, prepositions, conjunctions, pronouns and auxiliary verbs are removed as well as many general words. (VICK88) (BROO87).

TOME Searcher

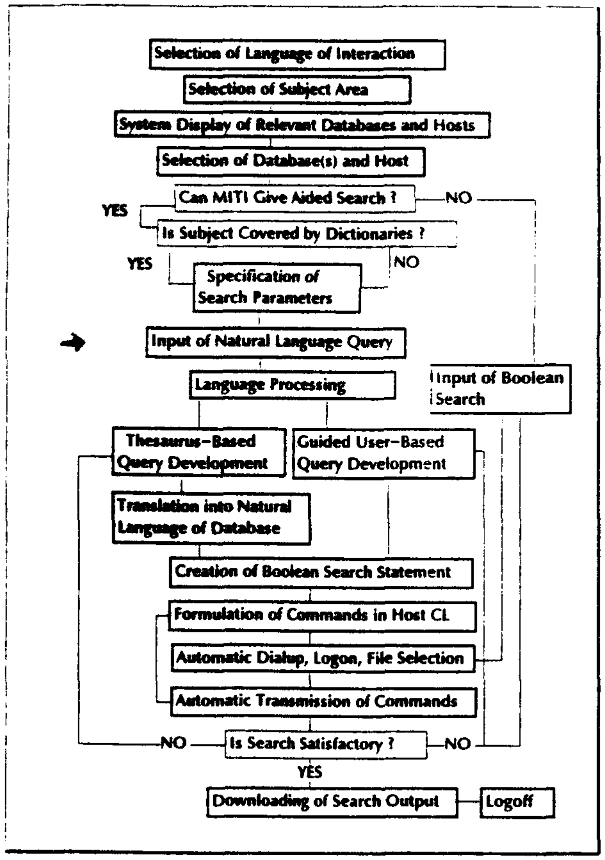
TOME Searcher is essentially a Natural Language system centered around the INSPEC Thesaurus. In 1987 it was decided to develop the PLEXUS system outside the University into an intelligent interface to online databases on hostcomputers. This product, TOME Searcher, was launched in the summer of 1988 by TOME Associates.
TOME Searcher was developed for professionals in electrical and electronic engineering, computer science and information technology (INSPEC).
The system has the functions of automatic dialup, logon, file selection and transmission of a search statement to the host computer, using the Common Command Language (CCL) of ESA.

Because semantic categories in electronic engineering etc are considerably different from those in the biological domain, a new semantic analysis of the terminology was needed.
TOME Searcher develops a first search strategy and automatically assesses the probable hit rate before going online, by consulting a thesaurus which contains the posting data from the databases covered (VICK90).
Because the terminology changes, the TOME Searcher Thesaurus has to be updated.

MITI

MITI is essentially a Natural Language system, centered around the use of the faceted ROOT Thesaurus of the British Standards Institute (VICK90).
The system will initially access four hosts : STN, ESA, Telesystemes and ECHO, but will be extendable to any host.

MITI – HIGH LEVEL DESCRIPTION


Selection of Language of Interaction
Selection of Subject Area
System Display of Relevant Databases and Hosts
Selection of Database(s) and Host
Can MITI Give Aided Search ?
YES
NO
Is Subject Covered by Dictionaries ?
YES
NO
Specification of Search Parameters
Input of Natural Language Query
Language Processing
Input of Boolean Search
Thesaurus-Based Query Development
Guided User-Based Query Development
Translation into Natural Language of Database
Creation of Boolean Search Statement
Formulation of Commands in Host CL
Automatic Dialup, Logon, File Selection
Automatic Transmission of Commands
NO
Is Search Satisfactory ?
NO
YES
Downloading of Search Output
Logoff

STINET

			NLP			NLP
HOST	DGIS	STILLAS	CCL	SEARCHME	ASTRO	OBES
		RIM		EASYNET	CCL	CONIT
BRS	+	+	+	+	+	
COMPUSERVE	+					
DARC	+					
DATASTAR				+		
DIALOG	+	+	+	+	+	+
ECHO	+					
ESA	+					
EUROOOEENNE DES DONNEES				+		
HAYSTACK	+					
INFOLINE	+					
LEXIS	+					
NEWSNET	+			+		
ORBIT	+		+	+		+
PERGAMON				+		
PERISCOPE/USNI	+					
PROFILE				+	+	
STN	+					
TELESCAN				+		
TELESYSTEMES/QUESTEL	+			+		
VU/TEXT				+	+	
WILSON LINE	+			+	+	
cindas	+					
dtic/drols	+		+			
elhill						+
matris	+					
mpo	+					
nasa	+		+			

Figure 2

The target domains of the current MITI will be : General Technology and Science, Environmental Issues (VICK90) MITI will develop an Intelligent Multilingual (English, French, German, Spanish) Interface which can be installed on a Personal Computer. It will enable Untrained Users to have access to different databases on a number of hosts in a uniform way. using Natural Language (CORD90)
MITI combines the best properties of the TOME Searcher and the IANI interface of the Scandinavian network (BOUM90).
In october 1991 the MITI project seemed to be withdrawn.

3.2.4 DIANEGUIDE/I'M GUIDE/MIM

The French firm ERLI (Etudes et Recherche Linguistique et Informatique) was founded in 1977 and now has some 70 employees working for government and private companies on Artificial Intelligence and Retrieval (Computational Linguistics). In 1990 France Telecom has taken shares in GSI-ERLI.
GSI-ERLI has developed several products such as :

- SAPHIR, a Natural Language Analyser, type ATN and comparable with INTELLECT, for interrogating factual databases, which is available for IBM mainframes working under MVS and VM/CMS (SQL and DB2). SAPHIR translates a question formulated in Natural Language for a relational database into SQL. The user does not have to be familiar either with the structure of the database or SQL.
- NLS, a Natural Language System to query the 2500 professional headings of the French Yellow Pages directory, which is available to 4 million End-Users with Minitel terminals (CLEM88).

GSI-ERLI has developed Natural Language database access software which uses the firm's propietary ALEXIS database management software. ERLI's intelligent front-end (sometimes refered to as ALEXDOC or "NLQP" (for natural language query processor) can be adapted to a variety of retrieval and database-systems, including OPACs. ERLI NLQP for bibliographic retrieval has been adopted as the retrieval software for access to the French online subject authority file RAMEAU. The RAMEAU file is maintained by the Bibliotheque Nationale and is available as public file to libraries via SUNIST, the national university network for STI (HILD89). RAMEAU is based on the Library of Congress Subject Headings (LCSH). RAMEAU contains some 100.000 Subject Headings, which are used in French libraries as the Common Subject Indexing language (JOUG89).

DIANEGUIDE

The European Community has given a contract to GSI-ERLI to develop a Natural Language Interface for the DIANEGUIDE database, which contains information in 9 languages about some 1500 European databases. Information is given about :

- the database (10 fields, such as name and subject)
- the host (11 fields, such as retrieval software)
- the producer (9 fields)

The subject of a database is defined by some 25 keywords and a summary.

The dictionaries which have been used were very detailed for finance, culture and sports, but less detailed for Justice and Geography. This could be solved by manual and intellectual control of the automatic indexing of the databases.

In the Natural Language access an untrained End-User can ask :

What are the databases dealing with Medicine? The answer is : 301. He then can choose to 1. See the results, 2. Narrow, 3. Broaden, or 4. Abandon

He will probably choose to Narrow by :
1. Language, 2. Database type
When he types 1 the following display will appear :
1. Language = English 200 hits
2. Language = German 100 hits
3. Language = French 1 hits

But suppose that an End-User wants to find information about the subject "Interactive Videodisc". He wants to type "Interactive Videodisc" and get a list of databases that have that information, and preferably the best.

When the the subject of the database is described by 25 keywords such as Medicine, he might get an answer Zero on his question.

But the NL Dianeguide has a Matching Motor (Moteur d'appariement MA) which will automatically broaden the query. First a predicate (AND/OR) tree is produced for predicates whose values index at least one record in the database. The Matching Motor then searches for all descriptors which are semantically close to the initial descriptor (automatic broadening of query).
But when Interactive Videodisc isn't in the database, there is no broadening.

I'M GUIDE

Information about European CD-ROM products has been added to the Dianeguide database, which then was combined with a Brokersguide. The new product was called I'M GUIDE. Databases are classified in 41 Categories such as Medicine, Education. 77 databases exist in the category Education. Although Natural Language Access is available, the I'M GUIDE still has no answer for the question Interactive Videodisc.

MIM

The knowledge that was developed during the Dianeguide project has been used by GSI for another IMPACT project, the Multilingual Interrogation Module (MIM). MIM has been developed for a database of 400 pages "A people's Europe", which contains 11 chapters with information about Equal treatment, Europe without frontiers etc. 40.000 words have been used for the English version, 60.000 for the French equivalent.
A multilingual dictionary has been used with 14.000 words : 8.500 UK/FR links, 6.500 UK/IT links and 4.900 FR/IT links.

CURRENT CONTENTS BASED DATABASE DIRECTORY (CCDD)

Because Natural Language is available in titles of Journal articles, a Database Directory might be based on the Current Contents of some 4000 European Scientific and Technical journals
(6 languages : Multilingual Contents) and 2000 from the US.

For each Journal-name details could be given about the databases that index & abstract articles from that journal. In that case an untrained End-User can ask "Interactive Videodisc", or "Intelligent Information Retrieval", find some relevant titles of articles and see in which journals and databases he can find more.

3.2.5. DGIS/STINET/CONIT

In 1980 Gladys Cotter wrote a report "Commercial database searching : a proposed additional DTIC user service" (COTT80). This report became a forerunner of the DGIS/STI Program, "STINET", which was described in detail in 1987 (COTT87). The STINET is based on several related development projects such as the Gateway, the Local Automation Model, the Directory of Resources, the Common Command Language, Post-Processing, an End-User Interface, an Expert Link and an Electronic Document System.
A DGIS/STINET bibliography was compiled in december 1988 (KUHN88).
The end goal of STINET is to bring these components together into a coherent and comprehensive whole and allow users to interact with information retrieval systems via an expansive Natural Language Interface.

THE DGIS GATEWAY

The DGIS Gateway is based on the Technology Information System (TIS) that was developed since 1975 by Victor Hampel at Lawrence Livermore National Laboratory (LLNL) under the sponsorship of the Department of Energy. Its Intelligent Gateway Processor (IGP) software was conceived in 1975 as a Table-driven interpreter for the creation of integrated Information Systems, the "Meta Machine" (HAMP79). The translation of dissimilar communication protocols in addition to the translation of commands and formats is carried out by the IGP's with an advanced version of the Network Access Machine (NAM) software which was developed by Rosenthal and Lucas for NBS. The NAM software was integrated in TIS in 1978 and completely rewritten in 1985 for TIS/IGP use (BURT86).

DGIS itself is a low-level AI-like system, which operates at DTIC on an integrated BSD UNIX and INGRES based software, called the IGP Toolbox. After a prototype DGIS had been running on a VAX 11/780 using UNIX, DGIS now resides on a Pyramid 98X Minicomputer at DTIC. In 1986 DGIS was evaluated by an professional information specialist. DGIS now connects to 21 database systems in native mode.

LOCAL AUTOMATION MODEL (LAM)

In the US there are 500 Defense libraries with over 10 employees. Some of them are run by contractors. Because DoD libraries felt the need to automate local information collections, DTIC in 1983 initiated the development of a Library Automation system, responsive to the local library management and networking needs of DoD libraries. The functional description of the Local Automation Model (LAM) was presented in October 1983 (HAMI83). Based on this study a prototype system was specified, the Integrated Bibliographic Information System (IBIS). IBIS would encourage wider participation of libraries in DTIC's Shared Bibliographic Input Network (SBIN) (COTT86).
A single command set should be available for local and central functions.
66 packages were identified, but no single system provided all the functions required. Because Integrated library systems supported most of the library functions and gateway systems supported external database access, but no package supported both, it was decided that the Intelligent Gateway Processor (IGP) of DGIS should form the gateway part of IBIS. 6 packages were identified for further selection. August 1985 the selected package was integrated with the gateway software (HART85). As refinement of the specifications progressed, DTIC was joined by the Library of Congress (LC) in its effort to develop a system suitable for use in Federal Libraries. In 1986 two versions of the LAM/IBIS prototype were tested (COTT87). September 1988 LC awarded a contract to SIRSI for thr Scientific and Technical Information Local Automation System (STILAS). An open contract between the Library of Congress and SIRSI enables Federal Government libraries to purchase STILAS though the FEDLINK program. STILAS is a turnkey-system, based on a combination of BRS/Search for retrieval and SIRSI's popular UNIX-based modular Unicorn Collection Management System, with modules for the online public access catalog, circulation control etc. The STILAS gateway permits simultaneous interaction with databases on Dialog or BRS, while also searching the local files. The universal access mode is made possible by the Retrieval Interface Manager (RIM). Essentially RIM is a translator, converting STILAS commands (in a format based upon BRS/Search) into the formats required for other systems (NEWT89).

DIRECTORY OF RESOURCES

A Directory of online databases was developed, which contains information on the content and scope of databases relevant to the interests of DoD. The Directory is making use of the INGRES RDBMS, which permits easy programming, unified use of native-mode or menu-driven mode for searching the Directory. But INGRES lacks support for large text fields and full-text retrieval. The Directory is subject-searchable, so that on entering the topic of interest, the user is provided a listing of appropriate databases (KUHN88) (KRUE90).

COMMON COMMAND LANGUAGE (CCL)

The DGIS Common Command Language (CCL) is a project to access the multiplicity of information systems with a standard command language. Because DGIS is a UNIX/C based system, the CCL began in 1986 with UNIX/C programming. Later on PROLOG was chosen to translate a CCL command into a command of a target database. Based on the design goals the CCL was structured as a knowledge-based system and evolved into a Common Command Language System (CCLS). DGIS CCL is based on the NISO/ANSI CCL (NISO87). DGIS CCL will gradually migrate from a structured language of NISO CCL to Natural Language. PROLOG will be coupled with a relational dbms with an SQL interface so that it can work with any RDBMS (KUHN88). The DGIS CCL is currently limited to single database access to major information systems DROLS/DTIC, NASA/RECON and 3 database vendors BRS, Dialog, ORBIT (TRAN89).

END-USER INTERFACE

Late 1984 Telebase started marketing of the Easynet system, which contained several components that should become available in DGIS/STINET. Because Easynet became very popular for End-User searching, the Easynet method was integrated in DGIS/STINET as the Menu-Aided Easy Seraching Through Relevant Options (SearchMAESTRO). It provides an easy way to search hundreds of Government and commercial databases without knowing the individual (native) search command language for each database. The prototype SearchMAESTRO was tested by approximately 30 users. The SearchMAESTRO was offered as a DTIC operational service in October 1987. Fall 1990 an User Needs Questionnaire was sent to all users. The most frequent cited reason for using SearchMAESTRO was that it eliminates the need for multiple accounts. The most useful features were SOS and SCAN. SOS gives immediate online assistance from search experts, and SCAN simultaneously searches several automatically-selected databases in a subject-aerea. It provides a list of databases with the number of hits.

In the interest of the more experienced searchers a menu-driven CCL interface was added to the SearchMAESTRO in february 1991. CCL allows for proximity searching and the building of sets. Extensive HELP screens and database-specific documentation is availbale while using the CCL feature. A TOTAL command tells the CCL user how much he has spent during the search session. Either ISO or NISO common commands can be used to search in the databases of the vendors BRS, Dialog, Profile, Wilson and VU/TEXT (GREE91).

NL QUERY BUILDING EXPERT SYSTEM

Because the SearchMAESTRO provides little in the way of sophisticated assistance for developing effective and comprehensive search strategies of the kind a human expert searcher could be expected to perform, in 1988 a limited number of DGIS users got access to the CONIT Advanced User Assistance on a Multics mainframe computer at MIT. CONIT, an acronym for "COnnector for Networked Information Transfer", was developed since 1981 by Richard Marcus and includes a Common Command Language and a menu-oriented interface mode (MARC88). CONIT is able to take a user's Natural Language phrase and to apply an all-fields keyword-stem approach to the automatic translation of the user's search request into a search strategy for any database and system. CONIT does not maintain Thesauri, but techniques for automatic phrase decomposition, common word exclusion and stemming. These techniques relate user's Natural Language topic expressions to both the free text and Thesaurus terms in the document's database records.

Example : "digitized document retrieval" is broken down into
"digit", "document" and "retriev"

F DIGIT? AND DOCUMENT? AND RETRIEV?
(Dialog)

Because the mainframe CONIT does not make use of modern interface techniques (e.g. windowing) and the explanations seem to have over-wordiness, the current version of CONIT has only limited possibilities for providing enhanced service for the DGIS community (MARC88)

The CONIT mainframe version was ported to a partial implementation of an "expert" version in UNIX/C in a Minicomputer environment.
Significant results were achieved in the implementation of the first phase of an algorithm that automatically ranks documents according to relevance models. Also developed was the design and partial implementation of an automatic search strategy narrowing selector based on user feedback or reasons for document irrelevance (MARC90).

Although these developments are interesting still much has to be done before QBES is operational and before End-Users have Natural Language access to the global database world.

POSTPROCESSING

The DGIS postprocessing utilities have been based on the bibliographic postprocessing capabilities of the TIS, which are described in a paper for the Online '82 Conference (HAMP82) and in (BURT85).

These capabilities included : Frequency analysis, merging of files, elimination of duplicates, cross correlation of fields and analysis of data field use.

DTIC now also wants to reach the End-user community, engineers with powerful workstations and spread-sheet software. Because user-needs of these End-users are different from the user-needs of bibliographic database searchers, DTIC has held an user-needs survey. In a paper of this AGARD-TIP meeting Huddy Haller reports about this user-needs survey.

Summary :
Natural language access is not yet available, but maybe the users of SearchMAESTRO and STILAS are quite pleased with their systems
and don't need NL access.

3.2.6 SPIRIT/EMIR/DIALECT

SPIRIT is a software package for full-text retrieval with reformulation. It is operational since 1981 and became popular since 1988. It is based on Natural Language Processing R&D by Christian Fluhr of INSTN and it is sold by SYSTEX. SPIRIT combines modular linguistic processing with statistical processing and accepts Natural Language queries. Text segments that are obtained as answers are ranked in descending order of relevance. Even a portion of a text can be used as a query.
In this case SPIRIT calculates the degree of semantic proximity between the query text and all other texts in the database and ranks selected documents in descending relevance order. The semantic proximity is calculated by using weights produced by a statistical model (FLUH89).

EMIR

European Multilingual Information Retrieval (EMIR) is a CEC ESPRIT project of Fluhr and SYSTEX, that will complete a feasibility study on automatic indexing of free-text and multilingual query of textual databases. It is based on SPIRIT and will use textual data concerning the building standards of nuclear plants or patent summaries.

DIALECT

DIALECT is an expert assistant for Information Retrieval that was developed by Bassano (BASS86). Fluhr and Bassano are discussing the possibility of combining EMIR and DIALECT.
By combining EMIR and DIALECT a powerful multilingual natural language system can be developed.

3.2.7 ALPHA DIDO

Hutton + Rustron Data Exchange Ltd, the publisher of the UK Defence Equipment Catalogue, is lead contractor and project manager of a two-year CEC IMPACT demonstration project for developing an online information service for the Construction Industry, which allows for Multilingual enquiries with particular relevance to the use of Standards.
The consortium includes The British Standards Institution, which will provide machine-readable data.

The project involves the use of an intelligent interface and domain knowledge models on the care of historic buildings, demolitions etc. The project is using Distributed Intelligence Data Operation (DIDO) as the method of operation.

The ENQ module is an interpretative intelligent interface with Natural Language features, operating on a reference engine (SYS). The SYS module embodies a concatenation (merging) of existing Thesauri such as :
BSI ROOT, TIT and ECCTIS.

SYS uses an interlexical system based on concept codes which operate multilingual.

4. NATURAL LANGUAGE AND THESAURUS AIDS

Vocabulary control is needed to alleviate the matching problems caused by the use of Natural Language in retrieval system queries and indexing.

The problems are of 3 types : morphological, syntactical and semantic. Thesaurus aids can be used to resolve semantic problems : control of homonyms or synonyms or other term equivalencies, and identification and classification of related, broader, and narrower terms. Most Thesauri incorporate both hierarchical and equivalence-type relationships. Users searching under a given term may need to find material indexed under an equivalent term, related terms, or narrower, more precise terms (HILD89)

A Thesaurus normally has the relations as shown in fig 4 (AITC90) (MILS90).

NATURAL LANGUAGE LEAD-IN TERMINOLOGY

A Thesaurus not only has synonyms but can also have Lead-in terminology. In that case instead of a non-preferred term (the lead-in term) a preferred term should be used. The non-preferred lead-in term is a Natural Language term which leads to the preferred Descriptor. This is done by USE and USE .. AND relationships.

MUNITIONS
USE Ammunition

ARMOR PIERCING PROJECTILES
USE Armor Piercing Ammunition
AND Projectiles

The USE ... AND relationship will also help the End-User to understand that he has to combine descriptors using the Boolean operator AND.

The more Lead-In terms a Thesaurus has, the lower the number of Descriptors can be. The more Natural Language Lead-In terms a Thesaurus has, the easier the Thesaurus will be for a End-Users. The Thesaurus then becomes an End-User Thesaurus (BATE86) or User Thesaurus (BATE89).

A Thesaurus without Natural Language Lead-In terminology is an Indexer Thesaurus, which can be used by trained Information Specialists.

An Indexer-Thesaurus should not be used by casual Users and Novice End-Users.

TOPICS

Instead of a Thesaurus, Reid suggests to use Topics, faceted semantic networks (or clusters of some 50 terms)(REID91). Generally, Topics can be compared with the Subject Categories that exist in several Thesauri such as INSPEC etc. The Topics are based on RUBRIC concept-trees (TONG85). RUBRIC (RUle-Based Retrieval of Information) uses a set of rules to make a heuristic decision tree containing word patterns. RUBRIC is able to weight terms according to their importance, as specified by the users.

A unique feature "modifier rules" gives RUBRIC synonym knowledge and helps to distinguish among multiple meanings of a term (HAWK88). RUBRIC combines knowledge-intensive techniques with efficient full-text retrieval and ranking strategies.

The RUBRIC approach seems particularly suited to users who are prepared to spend a lot of effort in constructing queries and would not be appropriate in a general environment (BELK87).

Expert users can build up librarie of retrieval topics, and Novice users can use them as building blocks to easily compose powerful queries (table V).

Acquiring the concepts and their qualitative and quantitative relationships require large amounts of effort. That's why the CONSTRUCTOR algorithm was developed by Tong to automatically generate relationships between concepts (building probabilistic networks from data). The CONSTRUCTOR system generates sparse networks (i.e. with few arcs) and can find subtle relationships that would take much effort to find manually. But there is still much work needed by a user to identify which Concepts are present (TONG90).

But if two people or groups of people construc . a Thesaurus in a given area, only 60 % of the index terms may be common in both Thesauri.

And if two scientists or engineers are asked to judge the relevance of a given set of documents to a given question, the area of agreement may not exceed 60 % (CLEV84).

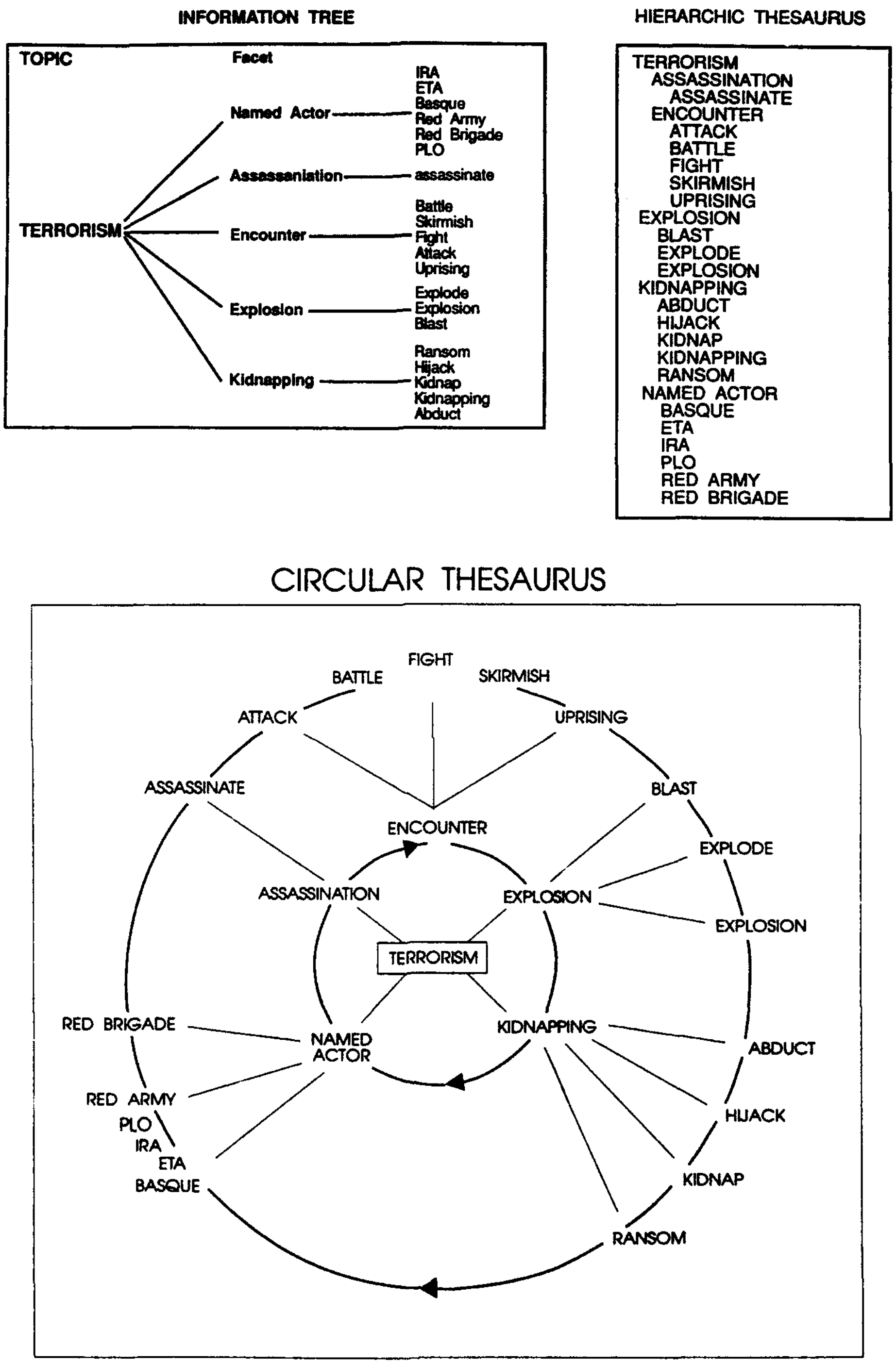
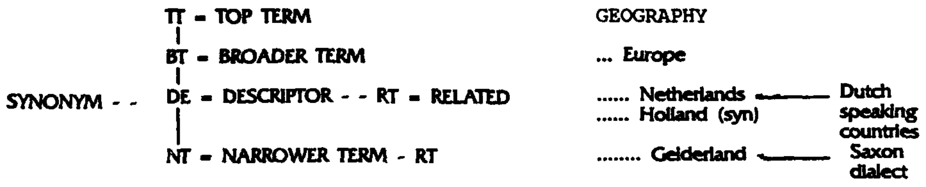
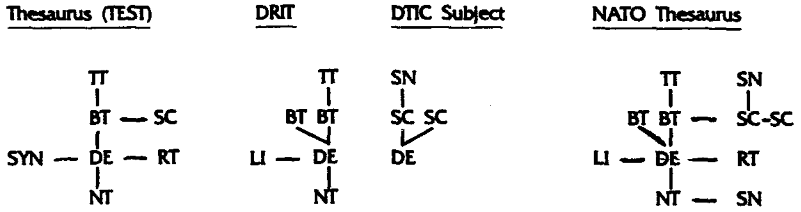

INFORMATION TREE
TOPIC
Facet
TERRORISM
Named Actor
IRA
ETA
Basque
Red Army
Red Brigade
PLO
Assassaniation
assassinate
Encounter
Battle
Skirmish
Fight
Attack
Uprising
Explosion
Explode
Explosion
Blast
Kidnapping
Ransom
Hijack
Kidnap
Kidnapping
Abduct
HIERARCHIC THESAURUS
TERRORISM
ASSASSINATION
ASSASSINATE
ENCOUNTER
ATTACK
BATTLE
FIGHT
SKIRMISH
UPRISING
EXPLOSION
BLAST
EXPLODE
EXPLOSION
KIDNAPPING
ABDUCT
HIJACK
KIDNAP
KIDNAPPING
RANSOM
NAMED ACTOR
BASQUE
ETA
IRA
PLO
RED ARMY
RED BRIGADE
CIRCULAR THESAURUS
FIGHT
BATTLE
SKIRMISH
ATTACK
UPRISING
ASSASSINATE
BLAST
ENCOUNTER
EXPLODE
ASSASSINATION
EXPLOSION
EXPLOSION
TERRORISM
RED BRIGADE
KIDNAPPING
ABDUCT
NAMED ACTOR
RED ARMY
PLO
HIJACK
IRA
ETA
BASQUE
KIDNAP
RANSOM

Figure 3

A Thesaurus normally has the following relations (ALTC90) :

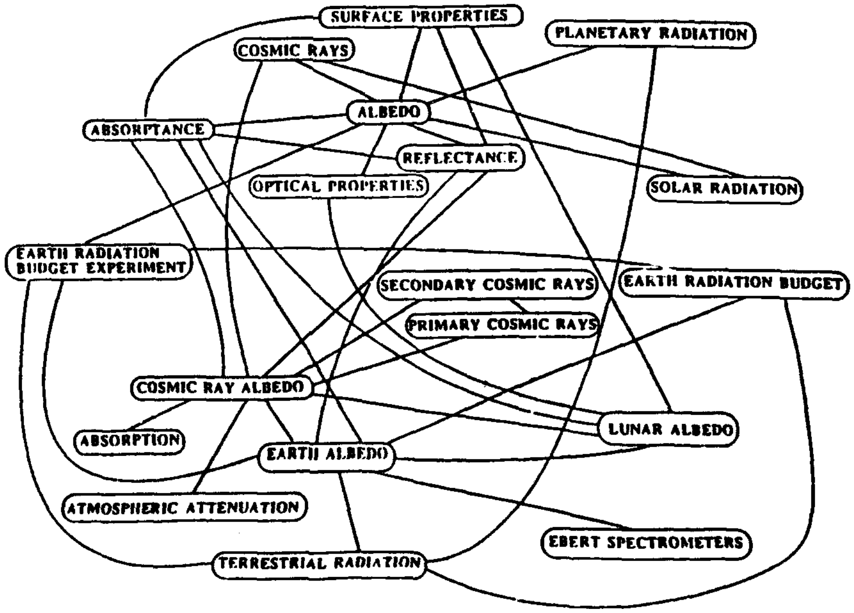


LEAD-IN TERMINOLOGY



SN = Scope Note (Definition)
SC = Subject Category (50 terms)

Polyhierarchy : 1 DE : 2 BT
1 BT : 2 SC



NASA related terms

Figure 4

Table V

NATO THESAURUS SUBJECT CATEGORIES

COSATI CODES														FIELDS
01	A	B	C12	D	E	F								AVIATION TECHNOLOGY
02	A	B	C	D	E	F								AGRICULTURE
03	A	B	C											ASTRONOMY AND ASTROPHYSICS
04	A	B												ATMOSPHERIC SCIENCES
05	A	B	C	D	E	F	G	H	I					BEHAVIORAL AND SOCIAL SCI
06	A	B	C	D	E	F	G	H	I	J	K	L	M	
	N	O												BIOLOGICAL AND MEDICAL SCIENCE
07	A	B	C	D	E	F								CHEMISTRY
08	A	B	C	D	E	F	G	H	I	J	K	L		EARTH SCIENCES AND OCEANOGRAPHY
09	A	B	C	D	E	F	G							ELECTROTECHN AND FLUIDICS
10	A	B	C	D										POWER PRODUCTION AND ENERGY
11	A	B	C	D	E	F2	G	H	I	J	K	L		MATERIALS
12	A	B	C	D	E	F	G	H	I					MATHEMATICS AND COMPUTER
13	A	B	C	D	E	F1	G	H	I	J1	K	L	M	MECHANICAL, INDUSTRIAL, CIVIL AND MARINE ENGIN
14	A	B	C	D	E									TEST EQUIPMENT, RESEARCH
15	A	B	C3	D	E	F7								MILITARY SCIENCES
16	A	B1	C	D3	E									GUIDED MISSILE TECHNOLOGY
17	A	B	C	D4	E2	F	G4	H	I	J	K			NAVIGATION DETECTION
18	A	B	C	D	E2	F	G	H	I	J				NUCLEAR SCIENCE AND TECHNOLOGY
19	A1	B	C	D	E	F	G	H1	I	J	K	L	M	ORDNANCE
20	A	B	C	D	E	F1	G	H	I	J	K	L	M	
	N	O												PHYSICS
21	A	B	C	D	E	F	G	H2	I2					PROPULSION ENGINES FUELS
22	A	B	C	D	E									SPACE TECHNOLOGY
23	A	B	C	D	E	F								BIOTECHNOLOGY
24	A	B	C	D	E	F	G							ENVIRONMENTAL POLLUTION
25	A	B	C	D	E									COMMUNICATIONS

COSATI STRUCTURE

SUBJECT CATEGORIES

SUBJECT CODES			SUBJECT HEADINGS
FIELDS			FIELDS
	GROUPS		GROUPS
01			AVIATION TECHNOLOGY
01	A		AERODYNAMICS
01	B		MILITARY AIRCRAFT OPERATIONS
01	C		AIRCRAFT
01	C	A	HELICOPTERS
01	C	B	BOMBERS
01	C	C	ATTACK AND FIGHTER AIRCRAFT
01	C	D	PATROL AND RECONNAISSANCE AIRCRAFT
01	C	L	RESEARCH AND EXPERIMENT AIRCRAFT

200 SUBJECT CATEGORIES	TEST (1967) NATO PCODATABASE WTI DATABASE
225 SUBJECT CATEGORIES	SIGLE DATABASE
250 SUBJECT CATEGORIES OF 50 DESCRIPTORS EACH = 12.500 DESCRIPTORS	DTIC NATO

250 TOPICS
OF 50 KEYWORDS

250 BAGS
OF 50 MARBLES

350 SUBJECT CATEGORIES	NTIS (SRIM)

Topics, Subject Categories or Thesauri Indexing and Searching can be difficult.

KNOWLEDGE GRAPHS

Knowledge graphs can represent many semantic structures found in Natural Language. Not only the structure of words can be represented, even the structure of sentences (HOED89). Integrated knowledge graphs can be used as an expert sysytem. The use of knowledge graphs may help a knowledge engineer to build an expert system by extracting knowledge from texts written by experts without having to consult them in person.
Knowledge graphs have been used for a medical diagnosis system MEDES and for describing the sulphur cycle in the Ems estuary (JAME90).

INDEX EXPRESSIONS

Index expressions have been used to show relations between concepts (BRUZ90,91).

4.1 MULTILINGUAL THESAURI

An Expert User normally can work with an Indexer Thesaurus without Lead-In Terminology, but sometimes even an Expert User needs help. That's the case when an US Expert User or an European End-User wants to search in European databases. Because of the language barrier they can't choose good Descriptors for searching in French or German databases. That's why they need a multilingual Thesaurus to search in European databases.

Multilingual Thesauri are available for multilingual databases such as : IRRD, EUDISED.

The French PASCAL database has a Multilingual Controlled Vocabulary called PASCAL LEXIQUE, which contains 80.000 controlled keywords in French, English and Spanish. French keywords have been used from the start of the database, but English keywords were only used since 1984.
UNESCO maintains a multilingual multidisciplinary SPINES Thesaurus which is being used for the SESAME database of CEC DG XVII on the Eurobases host.
The EURODICAUTOM database of the ECHO hostcomputer of CEC DG XIII contains the multidisciplinary multilingual terminology that is being used by CEC for automatic SYSTRAN translation.
The EUDISED Thesaurus is available in 9 European languages and is being used for the EUDISED database on the ESA hostcomputer. The EUDISED Thesaurus has some 3700 descriptors, with a graphical display. 2500 of the descriptors might be comparable or synonymous with 2500 descriptors from the ERIC Thesaurus and other Education-related Thesauri. When these Thesauri would become integrated, or linked, the EUDISED Thesaurus might become an multilingual Interface to ERIC and Education related databases.

```
          EUDISED            25
          French

                             250
English
EUDISED         2500 === 2500   ERIC

              + 1200 -- 25000
```

4.2 BILINGUAL NATO THESAURUS

In 1987 the NATO Standardization Group (NSG) felt the need for a NATO Standardization Information Base (NSIB), which should give information about ongoing Standardization activities. The NSG wanted to standardize the terminology by using a bilingual English/French Thesaurus. An AC/315 NSG Ad-Hoc Working Group advised in 1988 to use the DTIC Retrieval and Indexing Terminology (DRIT) as the baseline of the NATO Thesaurus (COTT89). In 1989 a Thesaurus Steering Group was formed. This group decided that the NATO Thesaurus should be a combination of the DRIT and the DTIC Subject Categorization Guide, which is used to distribute microfiches of reports (KRUE90).
By combining the DRIT with the DTIC Subject categories the structure of the NATO Thesaurus became similar to the structure of the TEST Thesaurus.
The NATO/DTIC Subject Categories are related to the SRIM Subject Categories of the NTIS database and to the Subject Categories of the NATO-PCO, SIGLE and WTI databases, which are also based on the COSATI Subject Categories of the TEST Thesaurus (TEST67).

Because the NATO Thesaurus allows for Lead-In terms, Natural Language Lead-In terms can be integrated in the NATO Thesaurus.

Because the NATO Thesaurus allows for Scope Notes for Subject Categories and Descriptors, the Scope Note facility can be used to integrate Natural Language Definitions of Descriptors or Subject Categories in the NATO Thesaurus. Definitions might be integrated at the Scope/Metadata level which exists in the NASA Thesaurus : 2500 terms (NASA89).

In this case the NATO Thesaurus will become a real End-User Thesaurus.
Unfortunately the AVOCON Thesaurus module which was used did not support the USE ... AND ... operator, so all DRIT USE ... AND terms had to be skipped. Generally, when software for Information Retrieval is chosen, the relationships of the Thesaurus module should be investigated.
An overview of possible relationships in Thesaurus software has been given by Milstead (MILS90).

NATO is investigating the possibility to introduce an USE ... + Scope Note instead of USE ... AND ... Because the AVOCON only supports a classification code of 4 digits, the DTIC Subject Code of 6 digits was transformed in an easy to remember 4 digit Subject Code : 15.01.02 became 15AB.

AMENDMENTS

Related terms were not available in DRIT, but should be available in the NATO Thesaurus. An experiment was made to extract Related terms from the NASA Thesaurus. NATO Terminology from Subject Category 01 Aerospace was transformed in NASA Descriptors using the NASA/DoD Switching language. The Related terms of these NASA descriptors were integrated and transformed back to NATO descriptors. The result was printed but could not be used because the NASA Thesaurus contains too many Related terms

The NATO Thesaurus was passed through a English spelling checker which identified Latin and American terminology. The American term has become a Lead-In term. European terminology can also be integrated as Lead-In term.

US:
ARMOR
USE Armour

NATO:
INSENSITIVE MUNITIONS
USE Insensitive explosions

EUROPE:
LOW VULNERABILITY AMMUNITION
USE Insensitive explosives

LINKING AND COMPATIBILITY

The system manager of the NATO Thesaurus has made a listing of the number of terms for each Subject Category, which has been combined with data about the number of AD-A reports in the NTIS Ondisc CD-ROM for 1990. This list shows for which Subject Categories new terminology should be added.

By linking/merging terminology from other Thesauri the compatibility with external databases can be promoted. Integrating and linking of multiple Thesauri is advocated by Alberico and Micco (ALBE90). Techniques for doing so are discussed by Mandel (MAND87).
In some cases a descriptor from an external database might become a NATO Lead-In term.

TRANSLATION

The translation of the NATO Thesaurus from English into French in 1990 has been a complicated process, because of the many partners that were involved. Tapes in different formats with different run-dates have given problems. But the problems have been solved and links between Enlish and French descriptors have been introduced. A toggle switch (function key) has been developed for online switching between English and French (online translation).

PRINTED EDITION

The edit copy of the English version of the NATO Thesaurus is available, as well as English NATO Thesaurus on diskette. Printed edition : January 92.

ANNEX

NL SEARCH STRATEGY :

INTELLIGENT INFORMATION RETRIEVAL

When online searching is done by intermediaries, the requester has to explain what he needs to the intermediary. During the interview the requester begins better to understand and describe the problem. When there is no interview the requester has to fill in a form. The requester is asked to explain the purpose of the request, and generally he is asked to mention a "good article" as an example.

When an untrained End-User wants to find information about a certain subject he sometimes does not exactly know what he wants and then he should also start with a good article as an example.

Because the title of an article, a report or a book is written in Natural Language, a search strategy in Natural Language generally can be based on (a part of) a title. So an untrained End-User can also search online in external databases, starting with Natural Language, without parsing, NLP.

Because this TIP specialist meeting is related to Artificial Intelligence, Intelligent Gateways and Information Retrieval, a good example of Natural Language is :

"Intelligent Information Retrieval".

There are several articles and papers with this title or with a title which contains this multiword subject. Maybe our untrained End-User has seen an article about IIR and wants to know more about it.

But : Intelligent Information Retrieval is not a Descriptor.

Which databases

Dialindex can be used to see which databases contain information about IIR.

Options are :

```
?sf all         = 352 files
?sf all social  = 42 files
?sf all infosci =  7 files only
```

```
s intelligent information retrieval
s intelligent(w)information(w)retrieval

s intelligent AND
  information AND retrieval
s intelligent AND
  information(w)retrieval

s intelligent(1N)
  information(1N)retrieval
```

The result is shown in table

The 18 hits in INSPEC have all been indexed with the identifier IIR, but 60 % of the 317 I+I+R hits in Information Science Abs are not-relevant.

The untrained End-User will start a Natural Language I I R search in INSPEC (b2) and print 10 records. He will probably be happy with 5 of them.

The Subject-Expert will start a clustersearch IwIwR in the databases INSPEC, COMPENDEX, ISA, LISA, ERIC, and NTIS (b 1, 2, 6, 8, 61, 202) and find some 140 records. After online removal of duplicates with (rd) he will print 100 records.

The Information Expert will build a strategy :

concept I : Information Retrieval

concept II : Artificial Intelligence, which contains Expert Systems, Knowledge Based Systems, Natural Language and the word Intelligent;

He will probably search in some 15 databases on Dialog and maybe several databases of other Hostcomputers.

	Costs
End-User	$ 10-20
Subject-Expert	$ 100-200
Information Specialist	$ 1000-2000

ESA Which databases?

Questindex Topic : Information Science

(sometimes with f)	I I R	IwIwR	I+IwR
INSPEC	36	36	384
Pascal 205	11	11	30
Pascal 204			14
ABI	9	9	79
NTIS	4	4	84
NASA	2	2	
	$ 0,3	$ 0,6	$ 0,6

Not in Topic Information Science :

COMPENDEX	225

Which keywords?

HYPERLINE shows which keywords are useful. The "get" command can be used to transfer the keywords to search

NTIS

HL intelligent information retrieval

information retrieval	IR
expert systems	ES
artificial intelligence	AI
databases	
user interfaces	

A good strategy would be

(IR) AND (ES OR AI)

INSPEC

HL intelligent information retrieval

information retrieval	IR
information retrieval systems	
expert systems	ES
user interfaces	
knowledge based systems	KBS

Inspec is more specific then NTIS.
A good strategy would be

(IR) AND (ES OR KBS)
HYPERLINE allows the End-User to start a search in Natural Language (IIR) and then find the relevant descriptors.

The End-User will start a very simple search in

		Hits	Prints
NTIS	: IR AND ES AND AI	23	10
INSPEC:	IR AND ES AND AI	48	10
	Total :		20
	Costs :		

The Subject-Expert starts a search in

NTIS	: IR AND (ES OR AI)	247	50
INSPEC:	IR AND (ES OR KBS)	368	
	: IR AND Intelligent : (I+IwR)	384	
	: IR AND (AI OR ES OR KBS)	471	
	: IR AND (AI OR ES OR KBS) (OR Intelligent)	601	50
ABI/INFORM			
	: (I+IwR)	79	10
COMPENDEX			
	: (I+IwR)	219	10
LISA	*: (I+IwR)*	289	10
	Total :		130
	Costs :		

The documents that are found by the Subject-Expert will be quite new

This is a good result for End-User and Subject-Specialist.

The Information Specialist will do it better and will make a ZOOM (table 8).

INSPEC also mentions Easynet (10), PLEXUS (5), RUBRIC (4), SAFIR (4), CODER (3), CONIT (3), EURISKO (2), KISIR (2), NORDINFO (2), SPIRIT (2)

When he has not found enough Descriptors he can navigate in INSPEC Thes

hl Artificial Intelligence : 11861

3816	Knowledge Engineering
17128	Expert Systems
4718	Learning Systems
3446	Natural Languages
7381	Neural Nets

```
hl Natural Languages       :   3446
        11861 Artificial Intelligence
         2062ComputationalLinguistics
         9836 User Interfaces
        15307 Database Management
              Systems
         8109 Information Retrieval

hl Expert Systems          :  17128
    11861 Artificial Intelligence
     4209 Decision Support Systems
    21738 Explanation
     1638 Knowledge Acquisition
     4903 Knowledge Based Systems

hl Knowledge Based Systems :   4903
    17128 Expert Systems
     3037 Knowledge Representation
```

The Information Specialist will build a search profile

Concept 1 : Information Retrieval

Concept 2 : Artificial Intelligence
Computational Linguistics
Expert Systems
Knowledge Based Systems
Natural Language
Intelligent

After saving the search profile, it will be executed using Clustersearch or Onesearch in all relevant databases on several Hostcomputers. Duplicates will be deleted online, and then some 1000 records will be printed.

The information specialist will also make a ZOOM on Authors in ESA (table 9). or a ZOOM on Journal names.

He can introduce the names of these authors in the Pascal databases and then combine all relevant authors in a set of over 100 hits and then make a ZOOM on descriptors, to find relevant French Descriptors.

He can also search in Pascal 204 and 205 with Search Strategy

Concept 1 : Information Retrieval

Concept 2 : Artificial Intelligence
Knowledge Based Systems
Expert Systems

90 hits

90 hits : ZOOM

```
90 Recherche Information
89 Information Retrieval
82 Intelligence Artificielle
81 Artificial Intelligence
54 Recuperation Informacion
51 Inteligencia Artificial
18 Information System
16 Base Donnee
16 Representation Connaissances
15 Knowledge representation
15 Langage Naturel
15 Representacion Conocimientos
14 Natural Language
14 Sistema Informacion
13 Sistema Experto
13 Systeme Expert
12 Expert System
```

After entering these Descriptors he can build a French language Search strategy:

Concept 1 :
Recherche Information
Concept 2 :
Intelligence Artificielle
Representation Connaissances
Systeme Expert
Base Connaissance
Langage Naturel

Which results in 183 hits, twice as much as with the English language Search Strategy.

TABLE 1 DIALOG Which databases? Dialindex : b 411 : Options :

```
?sf all         = 352 files
?sf allsocial   =  42 files
?sf all infosci =   7 files,

s intelligent information retrieval        I I R  (NATURAL LANGUAGE)
s intelligent(w)information(w)retrieval    IwIwR
s intelligent AND information AND retrieval  I+I+R
s intelligent AND information(w)retrieval  I+IwR
s intelligent(1N)information(1N)retrieval  ININR
```

	I I R	IwIwR	I+I+R	ININR	I+IwR
	all infosci	all social		all	all
INSPEC	18	60			385
LISA	0	15	295	20	289
COMPENDEX				23	219
Trade & Indust ASAP					158
SCISEARCH					141
Information Sci Ab	0	23	317		95
NTIS	0	6		5	85
ABI/INFORM					79
ERIC	0	12	58		50
Pascal					34
	I I R	IwIwR	I+I+R	ININR	I+IwR
Costs	$ 0,25	$ 1,82	$ 1,82		$ 5,50

TABLE 2 ZOOM

NTIS :		ABI/COMPENDEX	INSPEC
247 hits		304 hits	601 hits
	Information Retrieval Systems	199	130
249	Information Retrieval	87	761
201	Artificial Intelligence	97	200
68	Expert Systems	56	376
	Expert System		50
	User Interfaces		165
	User Interface		11
46	Natural Languages		53
	Hypermedia		23
	Natural Language Processing	3	19
	Online Searching	39	22
28	Computational Linguistics		20
	Fuzzy Set Theory		19
15	Knowledge Bases	6	24
12	Heuristic Methods		
	Intelligent Gateway	5	
	Intelligent Information Retr	4	10
12	Knowledge Based Systems	4	102
	Knowledge-Based Systems	4	
12	Knowledge Representation		44

BIBLIOGRAPHY

AGOS91 AGOSTI, M., GRADENIGO, G., MARCHETTI, Pier G.
Architecture and functions for a Conceptual Interface to very large Online bibliographic collections
In : RIAO 91

ALBE90 ALBERICO, Ralph., MICCO, Mary
Expert Systems for Reference and Information Retrieval
(Supplements to computers in libraries ; 10)
Westport, CT : Meckler 1990
ISBN Nr 0-88736-232-X

AITC90 AITCHISON, Jean., GILCHRIST, Alan
Thesaurus Construction
London : Aslib, 1990

BASS86 BASSANO, J.C.
DIALECT: an expert assistant for information retrieval
Canadian AI Conference, Montreal, May 1986

BATE90 BATES, Marcia J.
Where should the person stop and the Information Specialist search interface start
Information Processing & Management v 26 n 5 pp 575-91, 1990

BATE89 BATES, Marcia J.
Rethinking Subject Cataloging in the Online Environment
Library Resources and Technical Services, v33 n4 p400-12 Oct 1989

BATE88 BATES, Marcia J.
How to use controlled vocabularies more effectively in online searching
Online, v n p 45-56, November 1988

BATE87 BATES, Marcia J.
How to use information search tactics online
Online, v n 47-54, May 1987

BATE8o BATES, Marcia J.
Subject Access in Online Catalogs : A Design Model
Journal of the American Society for Information Science,
v 37 n 6 p 357-376, .. 1986

BATE86 BATES, Marcia J.
Terminological assistance for the online subject searcher
In : Proceedings of the 2nd Conference on Computer Interfaces and Intermediaries for Information Retrieval
Boston, May 28-31, 1986
NTIS Nr : AD-A174 000

BELK90 BELKIN, Nicholas J., MARCHETTI, Pier.G.
Determining the Functionality and Features of an Intelligent Interface to an Information Retrieval System
In : 13th ACM/SIGIR, Brussels, 1990

BELK87 BELKIN, Nicholas J., CROFT, Bruce W
Retrieval Techniques
In : Annual Review of Information Science and Technology
(ARIST) v 22, 1987, Martha Williams, Editor. Elsevier Science

BELK86 BELKIN, Nick
What does it mean for an Information System Interface to be Intelligent
In : Proceedings of the 2nd Conference on Computer Interfaces and Intermediaries for Iformation Retrieval
Boston, May 28-31, 1986
NTIS Nr : AD-A174 000

BELK86 BROOKS, Helen M., Daniels, P.J., BELKIN, Nicholas J.
Problem descriptions and user models :
developing an intelligent interface for document retrieval systems
INFORMATICS 8, pp 191-214

BENT91 BENTON, Peter M
The Multilingual Edge
Byte, March 1991, pp 124-132

BORG86 BORGMAN, Christine
Why are Online Catalogs Hard to Use?
Lessons learned from Information Retrieval Studies
Journal of the American Society for Information Science,
v 37 n 6 pp 387-400, November 1986

BORG84 BORGMAN, Christine
The user's mental model of an information retrieval system :
effects on performance. PhD Thesis
Stanford University, Stanford, California, 1984
Dissertation Abstracts Nr 8408258

BOUM90 BOUMANS, Jak
Stand van zaken Intelligente Interfacen
In : Lezingen Online Informatie Konferentie Nederland
20-21 Februari 1990, Rotterdam
den Haag, NBLC (VOGIN, NVB)
ISBN 90-6252-352 8

BOYD90 BOYD, Trevor, WARNE, Karen
End-User searching within Glaxo Group Research Ltd :
The Dialog Medical Connection
In : End-User searching :
The effective gateway to published information
London : Aslib, 1990
ISBN 0-85142-238-1

BRAN89 BRANDHORST, Ted
What are the possibilities for Coordinating Education Information
Databases?
Paper presented at the annual Meeting of the American Educational
Research Association, San Francisco CA, March 29, 1989
NTIS Nr : ED 310 779

BREM91 BREMNER, Joe
Electronic Information Services:
Particularities and Self-Regulation
(Contracts and Codes of Conduct
In : Intellectual Property Rights
AGARD Nr : LS-181

BROO86 BROOKS, Helen M
An intelligent interface for document retrieval systems :
developing the problem and retrieval component. PhD Thesis
London, the City University, Department of Information Science, 86

BROO85 BROOKS, Helen M., Daniels, P.J., BELKIN, N.J.
Problem descriptions and user models : developing an intelligent
interface for document retrieval systems
INFORMATICS 8, pp 191-214

BRUZ90 BRUZA, Peter D., WEIDE, Theo P. van der
Two Level Hypermedia - An improved architecture for Hypertext
Dept. of Information Systems, Nijmegen University,
Technical Report no 90-5, March 1990

BRUZ91 BRUZA, Peter D., WEIDE, Theo P. van der
The Modelling and Retieval of Documents using Index Expressions
Dept. of Information Systems, Nijmegen University,
Technical Report no 91-3, February 1991

BURT85 BURTON, Hilary D
Bibliographic Post-processing with the TIS Intelligent Gateway :
Analytical and Communication capabilities. Sep 85
NTIS Nr : DE85018153 = AD-A163 066

BURT84 BURTON, Hilary D.
Integration of Common Command Languages
In : First Conference on Computer Interfaces and Intermediaries for
Information Retrieval
Williamsburg, VA, October 3-6, 1984
NTIS Nr : AD-A167 700

BYSO90 BYSOUTH, Peter
Evaluating the use of several approaches to online literature
retrieval by research scientists.
In : End-User searching :
The effective gateway to published information
London : Aslib, 1990
ISBN 0-85142-238-1

CLEM88 CLEMENCIN, Gregoire
Querying the French Yellow pages :
Natural Language access to the Directory
Information Processing & Management v 24 n 6 pp 633-49, 1988
Shorter version was presented at RIAO 88

CLEV84 CLEVERDON, Cyril
Optimizing convenient online access to bibliographic databases
Information Services & Use v 4 n 1-2 p 37-47 1984

CORD90 CORDIS Database : RTD Projects
CORDIS Acronym : MITI

COTT89 COTTER, Gladys A., BLADOS, Walt R.
Terminology Strategies for International Information Exchange
Defense Applied Information Technology Center, Aug 1989
NTIS Nr : AD-A214 147

COTT89 COTTER, Gladys A.
Global Scientific and Technical Information Network
In : Proceedings Online Information 88, London 6-8 dec, Vol II
Oxford : Learned Information, 1988

COTT88 COTTER, Gladys A., KUHN, Allan D.
The DoD Gateway Information System (DGIS)
The Department of Defense Microcomputer User's Gateway to the World
Microcomputers for Information Management v 5 n 2 pp 73-92, June 88

COTT87 COTTER, Gladys
The Scientific and Technical Information Network (STINET) :
Foundation for Evolution
Defense Technical Information Center, Alexandria, Sept 1987
NTIS Nr AD-A 189 750

COTT86 COTTER, Gladys A., HARTT, Richard W.
Integrated Bibliographic Information System
Concept and Application for Resource Sharing in Special Libraries
DTIC, june 1986
NTIS Nr : AD-A174 151 (See also AD-A157 700, AD-A161 700

COTT80 COTTER, Gladys A.
Commercial database searching :
A proposed additional DTIC user service
DTIC Office of Information Systems and Technology, march 1980
NTIS Nr : AD-A181 104

COUS91 COUSINS, Jill., ROBINSON, Lesley.
The Online Manual
Oxford, Blackwell Publishers, 1991

CROF87 BELKIN, Nicholas J., CROFT, Bruce W
Retrieval Techniques
In : Annual Review of Information Science and Technology
(ARIST) V 22, 1987
Martha Williams (Editor)
ELsevier Science Publishers

DOSZ90 DOSZKOCS, Tamas E., REGGIA, J., XIA LIN
Connctionist models and Information Retrieval
In : Annual Review of Information Science and Technology (ARIST)
V 25, pp 209-262
ISSN 0066-4200

DOSZ86 DOSZKOCS, Tamas E
Natural language processing in Information Retrieval
Journal of the American Society for Information Science
v 37 n 4 pp 191-96, July 1986

DOSZ83 DOSZKOCS, Tamas E.
From research to application :
The CITE Natural Language Information Retrieval System
In : Proceedings R & D in Information Retrieval, Berlin, May 1982
Lecture Notes in Computer Science nr 146
Berlin : Springer Verlag, 1983

EFTH90 EFTHIMIADIS, Efthimis N.,
Online Searching Aids:
A revieuw of Front Ends, Gateways and other Interfaces
Journal of Documentation, V 47, Nr 3, p 218-262, September 1990

EMIR90 European Multilngual Information Retrieval
Esprit project 5312 Technical Annex, October 1990

ESSE91 ESSENS, Peter M.J.D., McCANN, C.A., HARTEVELT, M.A.
An Exploratory Study of the Interpretation of Logical Operators in Database Querying
TNO Institute for Perception, Soesterberg, Netherlands
TNO-report IZF 1991 B-2, March 1991

EVAN91 EVANS, David A., GINTHER-WEBSTER, K., HART, M.
Automatic Indexing using selective NLP and first-order Thesauri
In : Proceedings RIAO 91, April 2-5, Barcelona

FEIG88 FEIGENBAUM, Edward et al
The Rise of the Expert Company : How Visionary Companies are using
Artificial Intelligence to achieve Higher Productivity and Higher
Profits
New York : Times Books, 1988

FIDE88 FIDEL, Raya
Extracting knowledge for intermediary expert systems :
The selection of search keys. Report for NSF grant IST 85-09719
Seattle, WA : GLIS-University of Washington, 1988

FIDE86 FIDEL, Raya
Towards Expert Systems for the Selection of Search Keys
Journal of the American Society for Information Science
v 37 n 1 pp 37-44, 1986

FINI86 FININ, Thimothy W., JOSHI, Aravino K., WEBBER, Bonnie Lee
Natural Language Interactions with Artificial Experts
Proceedings of the IEEE, V 74, n 7, pp 921-38, July 1986

FLET90 FLETCHER, J.D.
Effectiveness and cost of an Interactive Videodisc Instruction in
Defense Training and Education
Institute for Defense Analyses, Alexandria, July 1990
NTIS Nr AD-A228 387

FLUH89 DEBILI, F., FLUHR, Christian., RADASOA, P.
About Reformulation in Full-text IRS
Information Processing & Management, v 25, nr 6, p 647-57

FRAN91 FRANTS,V.I., SHAPIRO, J.
Algorithm for Automatic Constructic of Query Formulations
in Boolean Form
Journal of the ASIS, v 42 nr 1 pp 16-26, january 1991

GREE91 GREEVY, Dug (Telebase Systems Inc)
Common Command Language available on SearchMAESTRO
DTIC Digest, v 91 n 2 p 4, april 1991

HAMI83 HAMILTON, W.P.
Local Automation Model : Conceptual Design Document
Logistics Managment Institute, ap.il 1983
NTIS Nr : AD-A144 383

HAMP85 HAMPEL, Victor E., GARNER, Bruce L., MATTHEWS, Jack R
Intelligent Gateway Processors as Integrators of
CAD/CAM Networks
American Society of Mechanical Engineers (ASME) Symp
Miami Beach, FL, November 21, 1985
NTIS Nr AD-A163 036

HAMP83 HAMPEL, Victor E., BAILEY, C., KAWIN, R.A., LANN, N.A.
TIS (Technology Information System): An Intelligent Gateway Computer
for Information and Modeling Networks - Overview
Lawrence Livermore National Lab., CA, Aug 83
NTIS Nr : AD-A135 916

HAMP81 HAMPEL, Victor E.
Fact Retrieval for the 1980's
Proceedings of AGARD TIP Specialist Meeting
Munich, West Germany, 1981
AGARD Nr CP 304

HART90 HARTLEY, R.J., 'EEN, E.M., LARGE, J.A., TEDD, L.A
Beyond Brolean Searching
In : Online Searching : Principles and Practice
London, Bowker-Saur, 1990
ISBN Nr : 0-408 02290-6

HART86 HARTER, S.P.
Online Information Retrieval - Concepts, Principles and Techniques
Orlando, Academic Press Inc, 259 pag, 1986

HART85 HARTT, R.W.
Local Automation Model : Software Benchmarking : Test Plan
DTIC/LMI, March 1985
NTIS Nr : AD-A 154 349

HART85 HARTT, R.W.
Local Automation Model : Program Specification
User Access for Cataloging and Retrieval
NTIS Nr : AD-A179 411

HAWK88a HAWKINS, Donald T.,
Applications of Artificial Intelligence (AI) and Expert Systems for
Online Searching
Online , v n p 31-43, january 1988

HAWK88b HAWKINS, Donald T., LEVY, Louise R., MONTGOMERY, K.L.
Knowledge Gateways : the Building Blocks
Information Processing & Management, v 24 n 4 pp 459-468, 1988

HILD89 HILDRETH, Charles R.
Intelligent Interfaces and Retrieval Methods
Advances in Library Information Technology, Issue Number 2
Library of Congress, Cataloging Distribution Service, Washington
ISBN 0-8444-0626-0

HILD89 HILDRETH, Charles R.
The online catalogue : developments and directions
London : Library Association, 1989
ISBN 0-85365-708-4

HOED89 HOEDE, Cornelis., WILLEMS, Mark
Knowledge graphs and Natural Language
University of Twenthe, Faculty of Applied Mathematics
Memorandum no 811, October 1989

INGW84 INGWERSEN, Peter
A cognitive view of three selected online search facilities
Online Review, v 8 n 5, p 465-492

ISO 88 INTERNATIONAL ORGANIZATION FOR STANDARDS
Documentation - Commands for Interactive Search techniques
ISO/DIS 8777

JACO90 JACOBS, Paul S., RAU, Lisa F.
SCISOR : Extracting Information from On-line News
Communications of the ACM, v 33, n 11, pp 88-97, November 1990

JAME90 JAMES, P
Knowledge Graphs
University of Twenthe, Faculty of Mathematics
Memorandum no 945, march 1991

JOUG89 JOUGUELET, Suzanne
Subject Access in Online Catalogs : Developments in France
Cataloging & Classification Quarterly, v 10, n 1-2, pp 213-224

KAHL86 STANFILL, Craig., KAHLE, Brewster
Parallel Free-text *search on* the Connection Machine system
Communications of the ACM, v 29, nr 12, pp 1229-1239, sept 1986

KAHL91 KAHLE, Brewster., MEDLAR, Art.
An Information System for Corporate Users :
Wide Area Information Servers
Online, pp 56-60, september 1991

KORY90 KORYCINSKI, C., NEWELL, Alan F
Natural-language processing and automatic indexing
The Indexer, v 17 n 1, pp 21-29, april 1990

KRAN89 KRANCH, Douglas A
The Development and Impact of an Global Information System
Information Technology and Libraries, pp 384-92, december 1989

KRUE90 KRUEGER, Jonathan
NATO Thesaurus Project
Control Data Corp. May 1990
NTIS Nr : AD-A222 700

KRUE90 KRUEGER, Jonathan
Referral Directory System Specification; Final rept
Control Data, Mar 1990
NTIS Nr : AD-A219 900

KUHN88 KUHN, Alan D
STINET & DGIS Reference Publications Bibliography
Defense Technical Information Center, December 1988
NTIS Nr AD-A203 926

KUHN88 KUHN, Alan D
DGIS : The development toward Artificial Intelligence and
Hypermedia in Common Command Language
Defense Applied Information Technology Center Hypermedia Lab, 1988
Presented at 12th Online Information, London 5-6 December 1988
NTIS Nr AD-A203674

KUHN88 KUHN, Allan D., COTTER, Gladys
DGIS : The Department of Defense Microcomputer User's Gateway
to the World.
Microcomputers for Information Management, v 5 n 2pp 73-92

LARG90 LARGE, J.A.
Software Developments
In : End-User Searching :
The effective gateway to published information
Edited by Peter T. Bysouth
London : Aslib, 1990
ISBN 0-85142-238-1

LEE 90 LEE, Newton S
InfoStation : A multimedia access system for library automation
The Electronic Library, v 8 n 6 pp 415-421, December 1990

LEIG89 LEIGH, William., PAZ, Noemi
The Use of SQL and Second Generation Database Management Systems
for Data Procesing and Information Retrieval in Libraries
Information Technology and Libraries, pp 400-07, december 1989

MAHO90 MAHON, B., MEINKOHN, F., LELLA, G.
TOOTSI : creating a toolkit for building user interfaces to business
information services
Online Review, v 14 n 6, pp 378-388

MAND87 MANDEL, Carol A.
Multiple Thesauri in online Library bibliographic systems
Washington : Libarary of Congress, Cataloging Distribution Service

MARC88 MARCUS, Richard S.
Experimental Evaluation of CONIT in DGIS Gateway Environment
Laboratory for Information and Decision Systems MIT, February 1988
Detense Applied Information Technology Center DAITC/TR-88/012
NTIS Nr : AD-A204 721

MARC90 MARCUS, Richard S.
Advanced Retrieval Assistance for the DGIS Gateway
Laboratory for Information and Decision Systems MIT, March 1990
US Air Force STINFO Contribution 90/1
NTIS Nr : AD-A221 592

MACG89 MAC GILLAVRY, W.E.
Relationele Trefwoordenvelden
Open, v 21, n 4, pp 134-36

MILS90 MILSTEAD, Jessica L.
Thesaurus software packages for Personal Computers
Database, v .. n .. pp 61-65, December 1990

MISC89 MISCHO, William H., MOORE, Amy
Enhanced access to periodical literature within an online catalogue
environment
In : The online catalogue
Developments and directions
Edited by Charles Hildreth
Library Association, London, 1989

MISC87 MISCHO, William H., LEE, Younghyoun
End-User searching of Bibliographic Databases
In : Annual Review of Information Science and Technology (ARIST)
Volume 22, pp 227-63, 1987.
Martha Williams, Editor
Published for the American Society for Information Science
(ASIS) by Elsevier Science Publishers

MITE89 MITEV, Nathalie N
Ease of interaction and retrieval in online catalogues :
contributions of human-computer interaction research
In : The Online Catalogue
Charles R. Hildreth
The Library Association, London, 1989

MITE85 MITEV, Nathalie N., WALKER, Stephen
Information retrieval aids in an online public access catalogue :
automatic intelligent search sequencing
INFORMATICS 8, pp 215-226

NASA89 NATIONAL AERONAUTICS and SPACE ADMINISTRATION
The NASA scientific and technical information system :
Its scope and coverage
NTIS Nr : N89-15779

NISO87 National Information Standards Organisation NISO
American National Standards Institute
Common Command Language for Online
Interactive Information Retrieval
ANSI Z39.58-1987

NEWT89 NEWTON, Barbara, JOURDAIN, Janet
The Weapons Laboratory Technical Library :
Automating with STILAS
In : Proceedings AGARD-TIP Specialist Meeting, Brussels Oct 17, 1989
AGARD CP-466

PORT80 PORTER
An algorithm for suffix stripping
Program v 14 n 3 pp 130-37, july 1980

REID91 REID, Clifford A
Enterprise-Wide Document Management Issues
Mountain View, CA : Verity Inc, May 1991

RICC91 RICCIUTI, Mike
Universal Database Access
Datamation, pp 30-35, november 1, 1991

RICH89 RICHARDS, Evelyn
The Data Deluge : Exotic Systems may hold Key to Future Access
Washington Post, v 112 n 293, Sect H, Sept 24, p 1, 1989

SCHA84 SCHANK, Roger C., CHILDERS, Peter G.
The Cognitive Computer :
On Language, Learning and Artificial Intelligence
Reading, MA : Addison-Wesley, 1984

SHOV85 SHOVAL, Peretz
Principles, Procedures and Rules in an Expert System for IR
Information Processing & Management, v 21, n 6, pp 475-87, 1985

SORM89 SORMUNEN, Eero
An analysis of online searching knowledge for intermediary systems
Espoo, VTT Information Service, july 1989
NTIS Nr : PB90-122532

SPAR90 COPESTAKE, Ann, SPARCK JONES, Karen
Natural language interfaces to databases
The Knowledge Engineering Review v 5 n 4 pp 225-49, 1990

STRI88 DE STRICKER, Ulla
A menu interface for Boolean logic
Information Services & Use v 8 n 1 pp 39-46, 1988

TICH89 TICHY, Walter F., ADAMS, Rolf L., HOLTER, Lars
NLH/E : A Natural Language Help System
In : Proceedings 11th Int. Conf. on Software Engineering pp 364-75

TONG90 FUNG, Robert M., CRAWFORD, Stuart L., APPELBAUM, Lee A
TONG, Richard M.
An Architecture for Probabilistic Concept-Based Information Retrieval
In : ACM/SIGIR 90

TONG85 McCUNE, B.P., TONG, Richard M., DEAN, J.S., SHAPIRO, D.G.
RUBRIC : a system for Rule-Based Information Retrieval
IEEE Transactions on Software Engineering SE-11 v 9 p 939-45, 1985

TRAN89 TRAN, Duc T., KUHN, Allan D., BIXBY, Randy L.,
DoD Gateway Information System (DGIS) Common Command Language :
A retrospective on the introduction of PROLOG as the development tool
DAITC Hypermedia laboratory, DTIC CCL report No 4, May 1989
NTIS Nr : AD-A211 941

VICK90 VICKERY, Brian C., VICKERY, Alina
Intelligence and Information Systems
Journal of Information Science Principles & Practices, v16n1 ,65-70

VICK75 VICKERY, Brian C
Classification and indexing in science. 3rd ed.
Butterworth 1975

VICK90 VICKERY, Alina
The role of Intelligent Online Interfaces to
bridge the communication gap
In : Bridging the Communication Gap
Proceedings AGARD TIP Specialist Meeting
Trondheim, 5-6 September 1990
AGARD Nr CP 487

VICK87 VICKERY, Alina., VICKERY, Brian C., BROOKS, Helen., ROBINSON, B
A reference and referral system using expert system techniques
Journal of Documentation, v 43 n 1 pp 1-23, march 1987

VICK87 VICKERY, Alina., BROOKS, Helen
PLEXUS - The Expert System for Referral
Information Processing & Management, v 23 n 2 pp 99-117, 1987

VICK87 VICKERY, Alina., BROOKS, Helen
Expert Systems and their Applications in LIS
Online Review, v 11 n 3 pp 149-65, June 1987

WALK89 WALKER, Stephen
The Okapi online catalogue research projects
In : The online catalogue
Charles Hildreth
The Library Association, London, 1989

WARN88 WARNER, Amy J.
Natural language processing in Information Retrieval
Bulletin of the American Society for Information Science
v 14 n 6 pp 18-19, Aug-sept 1988

WEIS87 BATES, M., WEISCHEDEL, R
Tutorial : Evaluating Natural Language Interfaces
Proceedings 25th Annual Meeting of the ACL
Stanford, CA, 1987
Cambridge, MA : Bolt, Beranek and Newman

WILL86 WILLIAMS, Martha E.
Transparent Information Systems through Gateways, Front Ends, Intermediaries and Interfaces
Journal of the American Society for Information Science
v 37 n4 pp 204-14, July 1986

ZBOR91 ZBORNIK, Stefan., GLASEN, Fabian
ESURS: ein System zur Ubersetzung von Retrievalsprachen
Nachrichten Dokumentation, v 21 n , p 271-81

Non Boolean Search Methods in Information Retrieval

C. J. van Rijsbergen
University of Glasgow
Department of Computing Science
Glasgow, G12 8RZ

Introduction

Information retrieval is a wide, often loosely-defined term but in these pages I shall be concerned only with automatic information retrieval systems. Automatic as opposed to manual and information as opposed to data or fact. Unfortunately the word information can be very misleading. In the context of information retrieval (IR), information, in the technical meaning given in Shannon's theory of communication, is not readily measured. Nevertheless it has become apparent that there is a notion of information fundamental to the information retrieval process that underlies our intuitions about what it is we attempt to retrieve (see Van Rijsbergen, 1989). In many cases one can adequately describe this kind of retrieval by simply substituting 'document' for 'information' where a document may be text, image etc.. This implies that the process is concerned with the identification of certain kinds of objects, viz. documents. It does not explain what is the basis of this identification process, and here it is that the notion of information plays a role. In the case of Boolean retrieval, an attempt is made to measure the extent to which the information is contained in a particular document by establishing whether a document satisfies the request. In the non-Boolean case, this process of satisfaction ha; a measure of uncertainty attached to it. For example, one may wish to express this uncertainty through a process of plausible inference.

All search strategies are based on comparison between the query and the stored documents. Sometimes this comparison is only achieved indirectly when the query is compared with clusters (or more precisely with the profiles representing the clusters). Or indeed, sometimes the comparison is based on a comparison with documents within a context or neighbourhood of a given document. Frequently the comparison is iterative in that a user provides feedback after a first comparison which will affect the next comparison.

The distinctions made between different kinds of *search strategies* can sometimes be understood by looking at the query language, that is, the language in which the information need is expressed. The nature of the query language often dictates the nature of the search strategy. For example, a query language which allows search statements to be expressed in terms of logical combinations of keywords normally dictates a Boolean search. This is a search which achieves its results by logical (rather than numerical) comparisons of the query with the documents.

Boolean search

A Boolean search strategy retrieves those documents which are 'true' for the query. This formulation only makes sense if the queries are expressed in terms of index terms (or keywords) and combined by the usual logical connectives AND, OR, and NOT. For example, if the query $Q = (K_1 \text{ AND } K_2)$ OR $(K_3 \text{ AND } (\text{NOT } K_4))$ then the Boolean search will retrieve all documents indexed by K_1 and K_2, as well as all documents indexed by K_3 which are *not* indexed by K_4.

An obvious way to implement the Boolean search is through the inverted file. We store a list for each keyword in the vocabulary, and in each list put the addresses (or numbers) of the documents containing that particular word. To satisfy a query we now perform the set operations, corresponding to the logical connectives, on the K_i-lists. For example, if

K_1 -list	:	D_1, D_2, D_3, D_4
K_2 -list	:	D_1, D_2
K_3 -list	:	D_1, D_2, D_3
K_4 -list	:	D_1

and $Q = (K_1 \text{ AND} K_2)$ OR $(K_3 \text{ AND } (\text{NOT } K_4))$

then to satisfy the $(K_1 \text{ AND } K_2)$ part we intersect the K_1 and K_2 lists, to satisfy the $(K_3 \text{ AND } (\text{NOT } K_4))$ part we subtract the K_4

list from the K_3 list. The OR is satisfied by now taking the union of the two sets of documents obtained for the parts. The result is the set $\{D_1, D_2, D_3\}$ which satisfies the query and each document in it is 'true' for the query.

A slight modification of the full Boolean search is one which only allows AND logic but takes account of the actual *number* of terms the query has in common with a document. This number has become known as the *co-ordination level.* The search strategy is often called *simple matching.* Because at any level we can have more than one document, the documents are said to be *partially* ranked by the co-ordination levels.

For the same example as before with the query $Q = K_1$ AND K_2 AND K_3 we obtain the following ranking:

Co-ordination level

3	D_1, D_2
2	D_3
1	D_4

In fact, simple matching may be viewed as using a primitive matching function. For each document D we calculate $|D \cap Q|$, that is the size of the overlap between D and Q, each represented as a set of keywords.

Another modification to simple Boolean matching derives from Fuzzy Logic. This work was pioneered by Zadeh (1965) and convincingly characterised by Bellman and Giertz (1973). In information retrieval several authors have worked applying the techniques, e.g. Bookstein (1980), Radecki (1979). There are fuzzy versions of the conventional set operations. If a document is in set A to degree a, and in set B to degree b, where a is greater than b, then it is in:

- the union of A and B to the degree a
- the intersection of A and B to the degree b
- the complement of A to degree $1-a$, and the complement of B to degree $1-b$.

Matching functions

Many of the more sophisticated search strategies are implemented by means of a matching function. This is a function similar to an association measure, but differing in that a matching function measures the association between a query and a document or cluster profile, whereas an association measure is applied to objects of the same kind. Mathematically the two functions have the same properties; they only differ in their interpretations.

There are many examples of matching functions in the literature. Perhaps the simplest is the one associated with the simple matching search strategy.

If M is the matching function, D the set of keywords representing the document, and Q the set representing the query, then:

$$M = \frac{2|D \cap Q|}{|D| + |Q|}$$

is another example of a matching function. It is, of course, the same as Dice's coefficient of a well known coefficient from the numerical taxonomy literature (Sneath and Sokal, 1973).

A popular one used by the SMART project (Salton, 1971), which they call cosine correlation, assumes that the document and query are represented as numerical vectors in t-space, that is $Q = (q_1, q_2, \ldots, q_t)$ and $D = (d_1, d_2, \ldots, d_t)$ where q_i and d_i are numerical weights associated with the keyword i. The cosine correlation is now simply

$$r = \frac{\sum_{i=1}^{t} q_i d_i}{\left(\sum_{i=1}^{t} (q_i)^2 \sum_{i=1}^{t} (d_i)^2 \right)^{\frac{1}{2}}}$$

or, in the notation for a vector space with a Euclidean norm,

$$r = \frac{(Q, D)}{\|Q\| \, \|D\|} = \text{cosine } \theta$$

where θ is the angle between vectors Q and D. The norms do not need to be the Euclidean norm; Salton (1989) has investigated a range of different norms.

Serial search

Although serial searches are acknowledged to be slow, they are frequently still used as parts of larger systems. They also provide a convenient demonstration of the use of matching functions. More importantly, they allow the specification of query and document representation without paying much attention to efficiency considerations. This means that once a matching function has been defined, consideration can be given to speeding up its execution using parallel architectures without sacrificing any of the complexity and accuracy of the representation mechanisms. There are now a number of parallel architectures (Rasmussen, 1991) used to increase the efficiency of retrieval engines.

Suppose there are N documents D_i in the system, then the serial search proceeds by calculating N values $M(Q, D_i)$ from which the set of documents to be retrieved is determined. There are two ways of doing this:

(1) the matching function is given a suitable threshold, retrieving the documents above the *threshold and* discarding the ones below. If T is the threshold, then the retrieved set B is the set $\{D_i | M(Q, D_i) > T\}$.

(2) the documents are ranked in increasing order of matching function value. A rank position R is chosen as cut-off and all documents below the rank are retrieved so that $B = \{D_i | r(i) < R\}$ where $r(i)$ is the rank position assigned to D_i. The hope in each case is that the relevant documents are contained in the retrieved set.

The main difficulty with this kind of search strategy is the specification of the threshold or cut-off. It will always be arbitrary since there is no way of telling in advance what value for each query will produce the best retrieval.

The advantage of a serial search is of course that the matching function can be made as complex as one wishes without any concern for inversion to speed up access. More recent research using parallel finegrained SIMD architectures (Stanfill and Kahke, 1986), uses what amounts to a serial search except that it is implemented in parallel. It allows the use of signatures derived from superimposed coding for the representation of documents. These signatures are matched in their entirety against a query. The best scoring documents are retrieved.

Probabilistic Retrieval

The basic instrument we have for trying to separate the relevant from the non-relevant documents is a matching function. The reasons for picking any particular matching function have never been made explicit. In fact, mostly they are based on intuitive argument in conjunction with Ockham's Razor. Simple probability theory can tell us what a matching function should look like and how it should be used. The arguments are mainly theoretical but, in my view, fairly conclusive (van Rijsbergen, 1979). The only remaining doubt is about the acceptability of the assumptions, which I shall briefly discuss. The data used to fix such a matching function are derived from the knowledge of the distribution of the index terms throughout the collection of documents or some subset of it. If it is defined on some *subset of documents* then this subset can be defined by a variety of techniques: sampling, clustering, or trial retrieval. The data thus gathered are used to set the values of certain parameters associated with the matching function. Clearly, should the data contain relevance information, then the process of defining the matching function can be iterated by some feedback mechanism similar to the one due to Rocchio described later in this paper. In this way, the parameters of the matching function can be 'learnt'. It is on matching functions derived from relevance information that we shall concentrate.

It will be assumed in the sequel that the documents are described by binary state attributes, that is, absence or presence of index terms. This is not a restriction on the theory; in principle the extension to arbitrary attributes can be worked out, although it is not clear that this would be worth doing (Osborne, 1975)

When we search a document collection, we attempt to retrieve relevant documents without retrieving non-relevant ones. Since we have no oracle which will tell us without fail which documents are relevant and which are non-relevant, we must use imperfect knowledge to guess for any given document whether it is relevant or non-relevant. Without going into the philosophical paradoxes associated with relevance, I shall assume that we can only guess at relevance through summary data about the document and its relationships with other documents. This is not an unreasonable assumption, particularly if one believes that the only way relevance can ultimately be decided is for the user to read the full text. Therefore, a sensible way of computing our guess is to try and estimate for any document its probability of relevance

$$P_Q \text{ (relevance/document)}$$

where the Q is meant to emphasise that it is for a specific query. It is not clear at all what kind of probability this is (see Good, 1950, for a delightful summary of different kinds), but if we are to make sense of it with a computer and the primitive data we have, it must surely be one based on frequency counts. Thus our probability of relevance is a statistical notion rather than a semantic one, but I believe that the degree of relevance computed on the basis of statistical analysis will tend to be very similar to one arrived at on semantic grounds. Just as a matching function attaches a numerical score to each document and will vary from document to document, so will the probability, for some it will be greater than for others and, of course, it will depend on the query. The variation between queries will be ignored for now, it only becomes important at the evaluation stage. So we will assume only one query has been submitted to the system and we are concerned with

$$P \text{ (relevance/document).}$$

Let us now assume (following Robertson, 1977) that:

(†) The *relevance* of a document to a request is independent of other documents in the collection.

With this assumption we can now state a principle, in terms of probability of relevance, which shows that probabilistic information can be used in an optimal manner in retrieval. Robertson attributes this principle to W. S Cooper although Maron in 1964 already claimed its optimality (Maron, 1965).

> *The probability ranking principle.* If a reference retrieval system's response to each request is a ranking of the documents in the collection in order of decreasing probability of relevance to the user who submitted the request, where the probabilities are estimated as accurately as possible on the basis of whatever data have been made available to the system for this purpose, the overall effectiveness of the system to its user will be the best that is obtainable on the basis of those data.

Of course, this principle raises many questions as to the acceptability of the assumptions. For example, the Cluster Hypothesis, that closely associated documents tend to be relevant to the same requests, explicitly assumes the contrary of assumption (†). Goffman (1964) too, in his work, has gone to some pains to make an explicit assumption of dependence. I quote: 'Thus, if a document *x* has been assessed as relevant to a query s, the relevance of the other documents in the file *X* may be affected since the value of the information conveyed by these documents may either increase or decrease as a result of the information conveyed by the document *x*'. Then there is the question of the way in which overall effectiveness is to be measured. Robertson in his paper shows the probability ranking principle to hold if we measure effectiveness in terms of Recall and Fallout.

But this is not the place to argue out these research questions. However, I do think it reasonable to adopt the principle as one upon which to construct a probabilistic retrieval model. One word of warning, the probability ranking principle can only be shown to be true for *one* query. It does not say that the performance over a range of queries will be optimised; to establish a result of this kind one would have to be specific about how one would average the performance across queries.

A detailed description of the probabilistic model can be found in Chapter 6 of van Rijsbergen (1979). There are several other good summaries e.g. Harper (1980), Salton & McGill (1983).

Cluster representatives

Before we can sensibly talk about search strategies applied to clustered document collections, we need to say a little about the methods used to represent clusters. Whereas in a serial search we need to be able to match queries with each document in the file, in a search of a clustered file we need to be able to match queries with clusters. For this purpose, clusters are represented by some kind of profile (a much overworked word), which here will be called a *cluster representative.* It attempts to summarise and characterise the cluster of documents.

A cluster representative should be such that an incoming query will be *diagnosed* into the cluster containing the documents *relevant* to the query. In other words, we expect the cluster representative to discriminate the relevant from the non-relevant documents when matched against any query. This is a tall order and, unfortunately, there is no theory enabling one to select the right kind of cluster representative (but see Croft, 1979). One can only proceed experimentally. There are a number of 'reasonable' ways of characterising clusters; it then remains a matter for experimental test to decide which of these is the most effective.

Let me first give an example of a very primitive cluster representative. If we assume that the clusters are derived from a cluster method based on a dissimilarity measure (see van Rijsbergen 1979), then we can represent each cluster at some level of dissimilarity by a graph (see *Figure 1*). Here A and B are two clusters. The nodes represent documents and the line between any two nodes indicates that their corresponding documents are less dissimilar than some specified level of dissimilarity. Now, one way of representing a cluster is to select a *typical* member from the cluster. A simple way of doing this is to find that document which is linked to the maximum number of other documents in the cluster. A suitable name for this kind of cluster representative is the *maximally linked document.* In the clusters A and B illustrated, there are pointers to the candidates. As one would expect in some cases, the representative is not unique. For example, in cluster B we have two candidates. To deal with this, one either makes an arbitrary choice or one maintains a list of cluster representatives for that cluster. The motivation leading to this particular choice of cluster representative is given in some detail in van Rijsbergen (1974a) but need not concern us here.

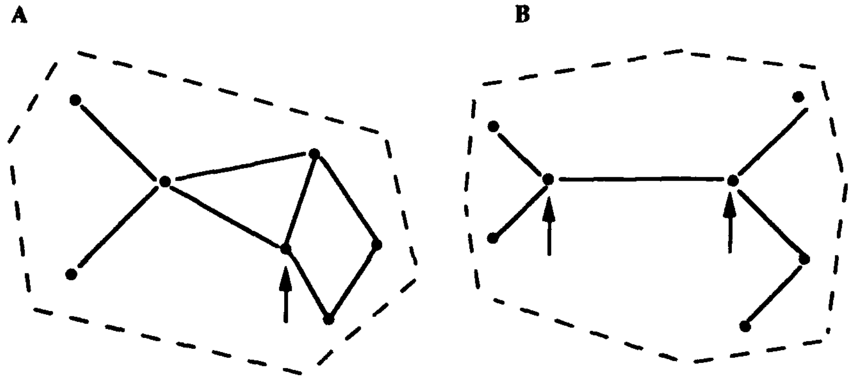


Figure 1. Examples of maximally linked documents as cluster representatives

Let us now look at other ways of representing clusters. We seek a method of representation which in some way 'averages' the descriptions of the members of the clusters. The method that immediately springs to mind is one in which one calculates the centroid (or centre of gravity) of the cluster. If $\{D_1, D_2, \ldots, D_n\}$ are the documents in the cluster and each D_i is represented by a numerical vector $(d_1, d_2, \ldots$

.., d_t) then the centroid C of the cluster is given by

$$C = \frac{1}{n} \sum_{i=1}^{n} \frac{D_i}{\| D_i \|}$$

where $\| D_i \|$ is usually the Euclidean norm, i.e.

$$\| D_i \| = \sqrt{d_1^2 + d_2^2 + \ldots + d_t^2}$$

More often than not the documents are not represented by numerical vectors but by binary vectors (or equivalently, sets of keywords). In that case, we can still use a centroid type of cluster representative but the normalisation is replaced with a process which thresholds the components of the sum ΣD_i. To be more precise, let D_i now be a binary vector, such that a 1 in the j th position indicates the presence of the j th keyword in the document and a 0 indicates the contrary. The cluster representative is now derived from the sum vector

$$S = \sum_{i=1}^{n} D_i$$

(remember n is the number of documents in the cluster) by the following procedure.

Let $C = (c_1, c_2, \ldots c_t)$ be the cluster representative and $[D_i]_j$ the jth component of the binary vector D_i, then two methods are:

$$(1) \quad c_j = \begin{cases} 1 \text{ if } \sum_{i=1}^{n} [D_i]_j > 1 \\ 0 \text{ otherwise} \end{cases}$$

or

$$(2) \quad c_j = \begin{cases} 1 \text{ if } \sum_{i=1}^{n} [D_i]_j > \log_2 n \\ 0 \text{ otherwise} \end{cases}$$

So, finally we obtain as a cluster representative a binary vector C. In both cases the intuition is that keywords occurring only once in the cluster should be ignored. In the second case we also normalise out the size n of the cluster.

There is some evidence to show that both these methods of representation are effective when used in conjunction with appropriate search strategies (see, for example, van Rijsbergen, 1974b, and Murray, 1972). Obviously there are further variations on obtaining cluster representatives but, as in the case of association measures, it seems unlikely that retrieval effectiveness will change very much by varying the cluster representatives. It is more likely that the way the data in the cluster representative is used by the search strategy will have a larger effect.

There is another theoretical way of looking at the construction of cluster representatives and that is through the notion of a maximal predictor for a cluster (Gower, 1974). Given that, as before, the documents D_i in a cluster are binary vectors then a binary cluster representative for this cluster is a predictor in the sense that each component (c_i) predicts that the most likely value of that attribute in the member documents. It is maximal if its correct predictions are as numerous as possible. If one assumes that each member of a cluster of documents $D_1, \ldots, D_n$ is equally likely, then the expected total number of incorrect predicted properties (or simply the expected total number of mismatches between cluster representative and member documents since everything in binary) is,

$$\sum_{i=1}^{n} \sum_{j=1}^{t} ([D_i]_j - c_j)^2$$

This can be rewritten as

$$\sum_{i=1}^{n} \sum_{j=1}^{t} ([D_i]_j - D_{\cdot j})^2 + n \sum_{j=1}^{t} ([D_{\cdot}]_j - c_j)^2 \qquad (*)$$

where

$$D_{\cdot j} = \frac{1}{n} \sum_{i=1}^{n} [D_i]_j$$

The expression (*) will be minimised, thus maximising the number of correct predictions, when $C = (c_1, \ldots, c_t)$ is chosen in such a way that

$$\sum_{j=1}^{t} ([D.]_j - c_j)^2$$

is a minimum. This is achieved by

$$(3) \qquad c_j = \begin{cases} 1 & \text{if } D._j > \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$$

So, in other words, a keyword will be assigned to a cluster representative if it occurs in more than half the member documents. This treats errors of prediction caused by absence or presence of keywords on an equal basis. Croft (1979) has shown that it is more reasonable to differentiate the two types of error in IR applications. He showed that to predict falsely 0 ($c_j = 0$) is more costly than to predict falsely a 1 ($c_j = 1$). Under this assumption the value of $^1/2$ appearing in (3) is replaced by a constant less than $^1/2$, its exact value being related to the relative importance attached to the two types of prediction error.

Although the main reason for constructing these cluster representatives is to lead a search strategy to *relevant* documents, it should be clear that they can also be used to guide a search to documents meeting some condition on the matching function. For example, we may want to retrieve all documents D_i which match Q better than T, i.e.

$$\{D_i \mid M(Q, D_i) > T\}$$

For more details about the evaluation of cluster representative (3) for this purpose the reader should consult the work of Yu and Luk (1977).

One major objection to most work on cluster representatives is that it treats the distribution of keywords in clusters as independent. This is not very realistic. Unfortunately, there does not appear to be any work to remedy the situation except that of Ardnaudov and Govorun (1977), and perhaps that of El-hamdouchi (1987).

Finally, it should be noted that cluster methods which proceed directly from document descriptions to the classification without first computing the intermediate dissimilarity coefficient, will need to make a choice of cluster representative *ab initio.* These cluster representatives are then 'improved' as the algorithm, adjusting the classification according to some objective function, steps through its iterations.

Cluster-based retrieval

Cluster-based retrieval has as its foundation the cluster hypothesis, which states that closely associated documents tend to be relevant to the same requests (van Rijsbergen and Sparck Jones, 1973). Clustering picks out closely associated documents and groups them together into one cluster. In van Rijsbergen (1979), Chapter 3, I discussed many ways of doing this; here I shall ignore the actual mechanism of generating the classification and concentrate on how it may be searched with the aim of retrieving relevant documents.

Suppose we have a hierarchic classification of documents then a simple search strategy goes as follows (refer to Figure 2 for details). The search starts at the root of the tree, node 0 in the example. It proceeds by evaluating a matching function at the nodes immediately descendant from node 0, in the example the nodes 1 and 2. This pattern repeats itself down the tree. The search is directed by a decision rule which, on the basis of comparing the values of a matching function at each stage, decides which node to expand further. Also, it is necessary to have a stopping rule which terminates the search and forces a retrieval. In Figure 2 the decision rule is: expand the node corresponding to the maximum value of the matching function achieved within a filial set. The stopping rule is: stop if the current maximum is less than the previous maximum. A few remarks about this strategy are in order:

(1) we assume that effective retrieval can be achieved by finding just one cluster;

(2) we assume that each cluster can be adequately represented by a cluster representative for the purpose of locating the cluster containing the relevant documents;

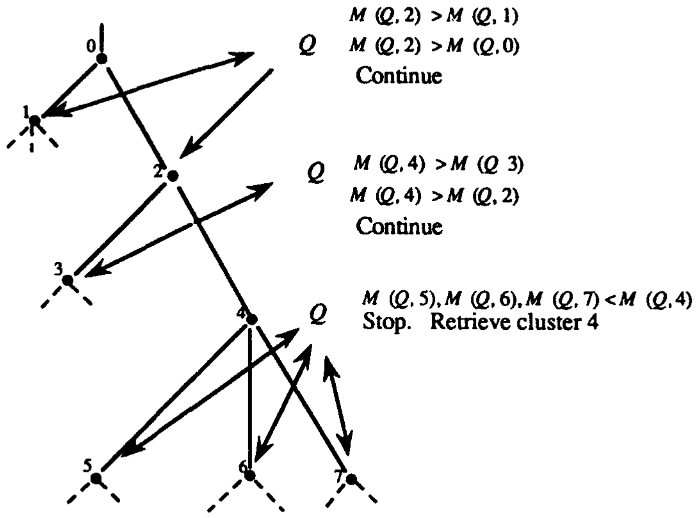


Figure 2. A search tree and the appropriate values of a matching function illustrating the action of a decision rule and a stopping rule.

(3) if the maximum of the matching function is not unique some special action, such as a look-ahead, will need to be taken;

(4) the search always terminates and will retrieve at least one document.

An immediate generalisation of this search is to allow the search to proceed down more than one branch of the tree so as to allow retrieval of more than one cluster. By necessity the decision rule and stopping rule will be slightly more complicated. The main difference being that provision must be made for *back-tracking*. This will occur when the search strategy estimates (based on the current value of the matching function) that further progress down a branch is a waste of time, at which point it may or may not retrieve the current cluster. The search then returns (back-tracks) to a previous branching point and takes an alternative branch down the tree.

The above strategies may be described as *top-down* searches. A *bottom-up* search is one which enters the tree at one of its terminal nodes, and proceeds in an upward direction towards the root of the tree. In this way it will pass through a sequence of nested clusters of increasing size. A decision rule is not required; we only need a stopping rule which could be simply a cut-off. A typical search would seek the largest cluster containing the document represented by the starting node and not exceeding the cut-off in size. Once this cluster is found, the set of documents in it is retrieved. To initiate the search in response to a request, it is necessary to know in advance one terminal node appropriate for that request. It is not unusual to find that a user will already know of a document relevant to his request and is seeking other documents similar to it. This 'source' document can thus be used to initiate a bottom-up search. For a systematic evaluation of bottom-up searches in terms of efficiency and effectiveness see Croft (1979).

If we now abandon the idea of having a multi-level clustering and accept a single-level clustering, we end up with the approach to document clustering which Salton and his co-workers have worked on extensively. The search strategy is in part a serial search. It proceeds by first finding the best (or nearest) cluster(s) and then looking within these. The second stage is achieved by doing a serial search of the documents in the selected cluster(s). The output is frequently a ranking of the documents so retrieved.

Interactive search formulation

A user confronted with an automatic retrieval system is unlikely to be able to express his information need in one go. He is more likely to want to indulge in a trial-and-error process in which he formulates his query in the light of what the system can tell him about his query. The kind of information that he is likely to want to use for the reformulation of his query is:

(1) the frequency of occurrence in the data base of his search terms;
(2) the number of documents likely to be retrieved by his query;
(3) alternative and related terms to be the ones used in his search;
(4) a small sample of the citations likely to be retrieved; and
(5) the terms used to index the citations in (4).

All this can be conveniently provided to a user during his search session by an interactive retrieval system. If he discovers that one of his search terms occurs very frequently, he may wish to make it more specific by consulting a hierarchic dictionary which will tell him what his options are. Similarly, if his query is likely to retrieve too many documents, he can make it more specific.

The sample of citations and their indexing will give him some idea of what kind of documents are likely to be retrieved and thus some idea of how effective his search terms have been in expressing his information need. He may modify his query in the light of this sample retrieval. This process, in which the user modifies his query based on actual search results, could be described as a form of *feedback*.

We now look at a mathematical approach to the use of feedback where the system *automatically* modifies the query.

Feedback

The word feedback is normally used to describe the mechanism by which a system can improve its performance on a task by taking account of past performance. In other words, a simple input-output system feeds back the information from the output so that this may be used to improve the performance on the next input. The notion of feedback is well established in biological and automatic control systems. It has been popularised by Norbert Wiener in his book *Cybernetics*. In information retrieval it has been used with considerable effect.

Consider now a retrieval strategy that has been implemented by means of a matching function M. Furthermore, let us suppose that both the query Q and document representatives D are t–dimensional vectors with real components where t is the number of index terms. Because it is my purpose to explain feedback, I will consider its applications to a serial search only.

It is the aim of every retrieval strategy to retrieve the relevant documents A and withhold the non-relevant documents $\bar{A}$. Unfortunately relevance is defined with respect to the user's *semantic* interpretation of his query. From the point of view of the retrieval system, his formulation of it may not be ideal. An ideal formulation would be one which retrieved only the relevant documents. In the case of a serial search the system will retrieve all D for which $M(Q,D) > T$ and not retrieve any D for which $M(Q,D) \leq T$, where T is a specified threshold. It so happens that in the case where M is the cosine correlation function, i.e.

$$M(Q,D) =$$

$$\frac{(Q,D)}{\|Q\| \; \|D\|} = \frac{1}{\|Q\| \; \|D\|} \times (q_1 d_1 + q_2 d_2 \ldots q_t d_t),$$

the decision procedure

$$M(Q,D) - T > 0$$

corresponds to a linear discriminant function used to linearly separate two sets A and $\bar{A}$ in R^t. Nilsson (1965) has discussed in great detail how functions such as this may be 'trained' by modifying the weights q_i to discriminate correctly between two categories. Let us suppose for the moment that A and $\bar{A}$ are known in advance, then the correct query formulation Q_0 would be one for which

$$M(Q_0,D) > T \qquad \text{whenever } D \in A$$

and

$$M(Q_0,D) \le T \qquad \text{whenever } D \in \bar{A}$$

The interesting thing is that starting with any Q we can adjust it iteratively using feedback information so that it will converge to Q_0. There is a theorem (Nilsson, 1965, page 81) which states that, providing Q_0 exists, there is an iterative procedure which will ensure that Q will converge to Q_0 in a *finite* number of steps.

The iterative procedure is called the *fixed-increment error correction* procedure.

It goes as follows:

$$
\begin{aligned}
Q_i &= Q_{i-1} + cD && \text{if} \quad M(Q_{i-1}, D) - T \le 0 \\
& && \text{and} \quad D \in A \\
Q_i &= Q_{i-1} - cD && \text{if} \quad M(Q_{i-1}, D) - T > 0 \\
& && \text{and} \quad D \in \bar{A}
\end{aligned}
$$

and no change made to Q_{i-1} if it diagnoses correctly. c is the correction increment, its value is arbitrary and is therefore usually set to unity. In practice it may be necessary to cycle through the set of documents several times before the correct set of weights are achieved, namely those which will separate A and $\bar{A}$ linearly (this is always providing a solution exists).

The situation in actual retrieval is not as simple. We do not know the sets A and $\bar{A}$ in advance, in fact A is the set we hope to retrieve. However, given a query formulation Q and the documents retrieved by it, we can ask the user to tell the system which of the documents retrieved were relevant and which were not. The system can then automatically modify Q so that at least it will be able to diagnose correctly those documents that the user has seen. The assumption is that this will improve retrieval on the next run by virtue of the fact that its performance is better on a sample.

Once again this is not the whole story. It is often difficult to fix the threshold*T* in advance so that instead documents are ranked in decreasing matching value on output. It is now more difficult to define what is meant by an ideal query formulation. Rocchio (1966) in his thesis defined the optimal query Q_0 as one which maximised:

$$\Phi = \frac{1}{|A|}\sum_{D \in A} M(Q,D) - \frac{1}{|\bar{A}|}\sum_{D \in \bar{A}} M(Q,D)$$

If M is taken to be the cosine function $(Q, D) / \|Q\| \, \|D\|$ then it is easy to show that Φ is maximised by

$$Q_0 = c\left(\frac{1}{|A|}\sum_{D \in A}\frac{D}{\|D\|} - \frac{1}{|\bar{A}|}\sum_{D \in \bar{A}}\frac{D}{\|D\|}\right)$$

where c is an arbitrary proportionality constant.

If the summations instead of being over A and $\bar{A}$ are now made over $A \cap B_i$ and $\bar{A} \cap B_i$ where B_i is the set of retrieved documents on the ith iteration, then we have a query formulation which is optimal for B_i a subset of the document collection. By analogy to the linear classifier used before, we now add this vector to the query formulation on the *i*th step to get:

$$Q_{i+1} = w_1 Q_i + w_2 \left[\frac{1}{|A \cap B_i|}\sum_{D \in A \cap B_i}\frac{D}{\|D\|} - \frac{1}{|\bar{A} \cap B_i|}\sum_{D \in \bar{A} \cap B_i}\frac{D}{\|D\|}\right]$$

where w_i and w_2 are weighting coefficients. Salton (1968), in fact, used a slightly modified version. The most important difference being that there is an option to generate Q_{i+1} from Q_i, or Q, the original query. The effect of all these adjustments may be summarised by saying that the query is automatically modified so that index terms in relevant retrieved documents are given more weight (promoted) and index terms in non-relevant documents are given less weight (demoted).

Experiments have shown that relevance feedback can be very effective. It is now one of the techniques that is frequently implemented in new operational systems.

Finally, a few comments about the technique of relevance feedback in general. It appears to me that its implementation on an operational basis may be more problematic. It is not clear how users are to assess the relevance, or non-relevance, of a document from such scanty evidence as citations. In an operational system it is easy to arrange for abstracts to be output but it is likely that a user will need to browse through the retrieved documents themselves to determine their relevance after which he may well wish to control the query adjustment himself or, at least, partially influence any automatic adjustments made.

References

ARNAUDOV, D.D. & GOVORUN, N.N. *Some Aspects of the File Organisation and Retrieval Strategy in Large Databases*, Joint Institute for Nuclear Research, Dubna (1977).

BELLMAN,R. & GIERTZ, M. 'On the analytic formalism of the theory of fuzzy sets', *Information Sciences*, **5**, 149-156 (1973).

BOOKSTEIN, A. 'Fuzzy requests', *Journal of the American Society for Information Science*, **31**, 240-247 (1980).

CROFT, W.B. *Organizing and Searching Large Files of Document Descriptions*, Ph.D. Thesis, University of Cambridge (1979).

EL-HAMDOUCHI, A. *Using inter-document relationships in information retrieval*, Ph.D. Thesis, University of Sheffield (1987).

GOFFMAN, W. 'A searching procedure for information retrieval', *Information Storage and Retrieval*, **2**, 294-304 (1977).

GOOD, I.J. *Probability and the Weighting of Evidence*, Charles Griffin and Co. Ltd., London (1950).

GOWER, J.C. 'Maximal predictive classification', *Biometrics*, **30**, 643-654 (1974).

HARPER, D.J. *Relevance feedback in document retrieval systems: an evaluation of probabilistic strategies*, Ph.D. Thesis, University of Cambridge (1980)

MARON, M.E. 'Mechanized documentation: The logic behind a probabilistic interpretation', In: *Statistical Association Methods for Mechanized Documentation* (Edited by Stevens *et al.*) National Bureau of Standards, Washington, 9-13 (1965).

MURRAY, D.M. 'Document retrieval based on clustered files', Ph.D. Thesis, Cornell University Report ISR-20 to National Science Foundation and to the National Library of Medicine (1972).

NILSSON, N.J. *Learning Machines - Foundations of Trainable Pattern Classifying Systems*, McGraw-Hill, New York (1965).

OSBORNE, M.L. 'A Modification of Veto Logic for a Committee of Threshold Logic Units and the Use of 2-class Classifiers for Function Estimation', Ph.D. Thesis, Oregon State University (1975).

RADECKI, T. 'Fuzzy-set theoretical approach to document retrieval', *Information Processing and Management*, **15**, 247-259 (1979).

RASMUSSEN, E.M. 'Parallel processing and information retrieval', Special Issue, *Information Processing and Management*, **27**, no. 4 (1991)

ROBERTSON, S.E. 'The probability ranking principle in IR', *Journal of Documentation*, **33**, 294-304 (1977).

ROCCHIO, J.J. 'Document retrieval systems - Optimization and evaluation', Ph.D. Thesis, Harvard University, Report ISR-10 to National Science Foundation, Harvard Computation Laboratory (1966).

SALTON, G. *Automatic Information Organisation and Retrieval*, McGraw-Hill, New York (1968).

SALTON, G. *The SMART Retrieval System - Experiment in Automatic Document Processing*, Prentice-Hall, Englewood Cliffs, New Jersey (1971).

SALTON, G. *Automatic Text Processing: The transformation, analysis, and retrieval of* *information by computer*, Addison Wesley, Reading, Mass. (1989).

SALTON, G. & McGILL, M.J. *Introduction to Modern Information Retrieval*, McGraw-Hill, New York (1983).

SNEATH, P.H.A. & SOKAL, R.R. *Numerical Taxonomy: The Principles and Practice of Numerical Classification*, W.H. Freeman and Company, San Francisco (1973).

STANFILL, C. & KAHLE, B. 'Parallel free-text search on the Connection Machine system', *Communiccations of the ACM*, **31**, 613-620 (1986).

van RIJSBERGEN, C.J. 'The best-match problem in document retrieval', *Communications of the ACM*, **17**, 648-649 (1974a).

van RIJSBERGEN, C.J. 'Further experiments with hierarchic clustering in document retrieval', *Information Storage and Retrieval*, **10**, 1-14 (1974b).

van RIJSBERGEN, C.J. 'Towards an information logic', *Proceedings of the 12th ACM SIGIR Conference*, 77-86 (1989).

van RIJSBERGEN, C.J. *Information Retrieval*, Second Edition, Butterworths, London (1979)

van RIJSBERGEN, C.J. & SPARCK JONES, K. 'A test for the separation of relevant and non-relevant documents in experimental retrieval collections', Journal of Documentation, **29**, 251-257 (1973).

YU, C.T. and LUK, W.S. 'Analysis of effectiveness of retrieval in clustered files', *Journal of the ACM*, **24**, 607-622 (1977).

ZADEH, L.A. 'Fuzzy sets', *Information and Control*, **8**, 338-353 (1965)

Hypertext and Hypermedia Systems in Information Retrieval

K. M. Kaye & A. D. Kuhn
NASA Scientific and Technical Information Program,
NASA Headquarters (Code NTT)
Washington, DC 20546
USA

Abstract: This paper opens with a brief history of hypertext and hypermedia in the context of information management during the "information age." Relevant terms are defined and the approach of the paper is explained. Linear and hypermedia information access methods are contrasted. A discussion of hyperprogramming in the handling of complex scientific and *technical information follows. A selection of innovative hypermedia systems is discussed.* An analysis of the Clinical Practice Library of Medicine NASA STI Program hypermedia application is presented. The paper concludes with a discussion of the NASA STI Program's future hypermedia project plans.

Introduction

The information management and media production environment have changed dramatically in the last ten years. The development of new communications systems, more powerful microcomputers, optical storage technologies, imaging and scanning technologies, and *animation and interactive video* technologies has dramatically altered the structure of society so that we now live in an information age. In today's society, the capability to retrieve, manage, and use information is of paramount importance. Information retrieval is concerned with the representation, storage and retrieval of documents or document surrogates. [CROU89] This technological revolution is changing the way we think of information retrieval and forcing an expanded definition of familiar terms such as "document". At one time a document meant a formal or legal paper such as a hardcopy technical *report. Now information pundits are adopting a broader* definition of document in saying that a document is "recorded information structured for human comprehension." [LEVI91] This paper approaches information retrieval with the broadest perspective possible in including documents in both familiar and still emerging forms.

This intensifying technological revolution has added impetus to the acceptance of new methodologies for information and knowledge retrieval and management. A *methodology to be addressed here is hypertext.* Ted Nelson coined the term hypertext in 1965 to describe a system of computer-supported, nonsequential information processing. Early hypertext systems developed on mainframe computers made it possible for users to create and explore information interactively. The central concept was the ability to create computer-supported links or cross-references permitting rapid, easy movement between related information. The ability of the user to control his path through the information and annotate or add to the information were also key concepts. Hypertext is actually a subset of a larger technology called *hypermedia* that is now available on computers ranging from mainframes to micros. Hypermedia extends the hypertext concept to link not only textual material, but all forms of material that may be digitally encoded for storage and retrieval through computer-based systems such as images, sound, graphics, and video [CHEN90]. To acknowledge the *multiple* forms of information in such systems, the term multimedia has been applied to hypermedia systems. In the literature today, one often sees the terms hypermedia and multimedia used interchangeably. However, whereas a hypermedia system can be correctly described as a multimedia system, a multimedia system is not necessarily a hypermedia system. The mere inclusion of multiple forms of information linked to each other is not enough to be a hypermedia system. Only when users can interactively take control of a set of dynamic links among units of information, can the system be correctly referred to as a hypermedia system.

Linear Access vs. Hypermedia Trails

Reading is fundamentally linear. Words are grouped together to form sentences, *sentences to form paragraphs,* paragraphs to form documents. Each has a beginning and is read through to the end. If the reader is searching for a specific piece of information contained in the document, the document is likely to be skimmed or perused linearly until the information is found. This is true of other linear mediums such as videotapes as well. With large information oriented documents such as texts, the reader may use an index to locate the unit containing the particular reference. The unit is then read linearly until the information is found. Even though the user can turn *directly to the location of the information, the act of* extraction is still linear. The document itself was designed to be accessed with a clear path through the information from beginning to end. Although the analogy is strained, this linear information access method can be viewed as corresponding to traditional information access methods as applied to such text products when reproduced in computer usable formats [BELL87]. In contrast, hypertext systems may provide the user with an initial linear access method, but at any given location in the information, the user has the option of selecting one to many further references. *In some svstems, a view of both* the incoming and outgoing references is available. As mentioned above, hypermedia systems allow these references to consist of any recorded information structured for human comprehension that can be accessed via a computer. Thus, with such systems the end user can pursue data references by following a self-selected trail or combination of trails through the data [BELL87].

Hyperprogramming and STI

Hyperprogramming is the process of creating *hypertext or* hypermedia applications. Although the inclusion of the root word "program" is used, professional programmers or software engineers are NOT required to create hypermedia applications. The newest hypermedia authoring systems have been designed to put hypermedia authoring into the hands of end users so that they can bring their ideas to life without having to master computer programming to do so.

Hypermedia is often presented as a new medium (akin to the invention of paper) that has tremendous potential to transform society. As a communications medium, hypermedia merges the three separate technologies of motion pictures, publishing, and computing with expertise drawn from all three disciplines to develop authoring systems designed to be accessible to any computer user.

Just as other new media have required a technological infrastructure before becoming widespread, hypermedia's popularity has required the availability of powerful desktop microcomputers with considerable storage available to handle large webs of multimedia nodes and links [WARR90]. Additionally, the recent emergence of systems that automatically generate hypertext documents from linear text documents has provided renewed interest in hypertext and hypermedia from organizations that originally rejected the technology as too labor-intensive.

Scientific and technical information (STI) differs from conventional and non-scientific information in two important ways. First, the syntactic and semantic elements found in the text are subject-area specific and consequently there is only a limited amount of overlap with those pieces of text which are also subject-area specific but whose subjects of discussion are different. Secondly, STI is, in general, rich in pragmatic content. The pragmatic aspects relevant to STI include items such as references to work by other authors, figures, tabular data, charts, mathematical expressions, and even videotapes of tests, equipment launches, *etc.* [MARS91]. Consider the following typical excerpt of scientific and technical information:

> *The science needs, as expressed at the High Resolution High Frame Rate Video Workshop held May 11-12, 1988 at NASA's Lewis Research Center (see Ref. 7), were largely validated by our review and analysis of those needs. Table 1 contains a summary of some of the greatest imaging needs. To be able to serve this domain of imaging and its attended data rates requires significant computer, recording, power, space, and weight resources (see Ref. 8).* [HAND90]

The typical user might be interested in looking at the video tapes of the workshop, examining the contents of table 1 and possibly manipulating them, and obtaining additional details on the article cited in *Reference 8*. All of these pieces of information could be accessed within a hypermedia system by activating a link. The ease of access stands in sharp contrast to juggling a cited article, a hardcopy table of data, a formula and/or software to manipulate the data, and a videorecorder to play the videotapes. Additionally, the additional complexity added by the data table and other pieces of information could be concealed or revealed to facilitate navigation through complex levels of abstraction.

The aerospace and defense community is primarily a community of scientists and engineers that are notorious for depending upon informal communication for information transfer. Hypermedia can improve upon the informal communication channels used for information transfer. The collaboration that now takes place via electronic mail and computer conferences can be enhanced by hypertext or hypermedia mechanisms. Communications networks that span national boundaries provide the connectivity necessary to allow multi-national participation in such collaborations with information nodes annotated by any participant at any time. As data compression and networking technologies advance, we will see hypermedia nodes and annotations in addition to text. Asynchronous annotation of information nodes is especially practical for collaborations across time zones where participants are not all likely to be awake at the same time. During the interim, computer-aided systems (like Systran) can be used to address the language barriers that exist. Thus, hypermedia technology can be used to enable collaboration despite barriers of distance, time, or language.

Within the aerospace and defense community, complex, long-term projects necessitate the creation of knowledge bases that transcend the involvement spans of individual personnel. NASA scientists and engineers have particularly mentioned their need to record the expertise of key individuals so that it will not be lost when they retire or leave. Hypermedia-based associative memories are ideal solutions to these problems since all of the rich pragmatic content typical of STI can be included. Hypermedia-based associative memories can recall information even when queries are incomplete or garbled, can store data in a distributed fashion, can detect similarities between new inputs and previously stored patterns, and do not degrade appreciably in performance if some of the memory's components are damaged--all useful characteristics for a distributed, shared organizational knowledge base [WARR90]. Hypermedia associative memories form the basis for the performance support systems that are just beginning to be seen in Government and industry.

The main value-added of hypermedia systems to the STI community lies in the ability of hypermedia to handle the full spectrum of STI's pragmatic content from data manipulation to video display. The presentation of standard textual information only just begins to take advantage of hypermedia's strengths and suffers the disadvantages associated with forcing users to read large amounts of text from today's computer screens [NIEL90].

In summary, hyperprogramming is well suited to STI. Advocates of hypermedia pose the following arguments for why hypermedia constitutes a major advance over other media:

o The associative, nonlinear nature of hypermedia mirrors the structure of human long-term memory, empowering both intelligence and coordination through intercommunication.

o The capability of hypermedia to reveal and conceal the complexity of content lessens the cognitive load on users of this medium, thereby enhancing their ability to assimilate and manipulate ideas.

o The structure of hypermedia facilitates capturing and communicating knowledge, as opposed to mere data.

o Hypermedic's architecture enables distributed, coordinated interaction, a vital component of teamwork, organizational memory, and other "group mind" phenomena [WARR91].

Although some would oppose the above claims, the unique characteristics of STI and the STI community serve to make this area particularly liable to benefit from greater use of hypermedia technology.

What's Happening Now?

There are a multitude of hypermedia developments going on in the transition from traditional linear information retrieval to actual information viewing. This is happening in what has been termed the *multidimensional information space* [SEPEC90]. We have seen from the above how this is changing the way we store, retrieve, and use information. Hyper-branching applications are being experimented with throughout the whole of the government, academia, and private industry. The following are but a few examples representative of what is taking place.

The Experiment Documentation Information System (EDIS. EDIS is being developed by Houston Applied Logic, Houston, Texas, for the NASA Life Sciences Project Division at NASA Johnson Space Center, Houston, Texas. It is a system designed to produce and control the Life Sciences Experiment Document (ED) containing large amounts of text in combination with tables and graphs of mathematical and scientific data, making use of hypertext concepts through Macintosh HyperCard. The ED defines all functional objectives, inflight equipment, consumables, measurements, ground support, and test sessions, along with the expected results of the experiments. The ED consists of 16 chapters plus appendices. There is a fixed, or boilerplate text in some sections that applies to any Life Sciences experiment and reference table formats concerning experiment-specific text and mathematical/scientific data. Other sections contain experiment data tailored for each experiment. The EDIS is foreseen as being the first step in the automation of the process required for defining complete packages of Life Sciences experiments for the Shuttle missions [MOOR90].

Life Sciences Interactive Information Recall (LSIIR). This is a study in hypermedia applications, being done by GE Government Services, Houston, Texas, for the Life Sciences Project Division, Johnson Space Center. LSIIR, through interactive media technologies, provides online information aids as a "job performance assistance." The technologies are integrated into a computer desktop workstation environment with which mission or payload specialist, the scientist, the engineer, and support or administrative people are familiar. The LSIIR is foreseen as providing assistance in Life Sciences Project missions and activities such as development and testing, science monitoring, technical lab activities, and mission testing. The system make uses Mac SEs running integrated applications of HyperCard, MacRecorder Sound System, MacDraw, MacPaint, Canvas, and MacroMind Director. MacroMind Director enhances graphics display and animation. Clip art and scanned photos are part of the system's information base. The system serves as a "trainer" or simulator. It provides the user with different sets of information to change variables during an exercise, or make alterations to procedures and configurations. LSIIR has passed its proof of concept stage, and is envisioned as an online system for electronic documentation and information, and electronic training and review in all areas of the NASA Life Sciences Project activity [CHRIS90].

Decision Support System Shell. The Carroll School of Management, Boston College, Chestnut Hill, Massachusetts, is constructing a decision support system (DSS) on a Macintosh that can support applications in a variety of fields such as engineering, manufacturing, and finance. The shell provides for a hypertext-style interface for navigating among DSS application models, data, and reports. They enhanced the traditional notion of manual, predefined hypertext links by allowing for hypertext connections to be built "on the fly." The 'erm "generalized hypertext" is applied, in the sense of networking links within a domain of multiple documents. Generalized hypertext is a logic-based technique for automating hypertext within a knowledge-based decision support environment. Value is added by providing the hypertext-style interface to the DSS application without an author having to create any nodes or links, while at the same time allowing for adding comments and other annotations. An earlier version of the current development, called MAX, was done for the U.S. Coast Guard [BIEB90].

Knowledge Base Browser (KBB). Currently under development at the NASA Johnson Space Center is a hypermedia system for browsing CLIPS knowledge bases. CLIPS is C Language Integrated Production System, an expert system shell used in this case to create knowledge base expert systems of rules that control the processes of the Onboard Navigation (ONAV) flight control position at the Mission Control Center (MCC). These expert systems will support the ascent, rendezvous, and deorbit/landing phases of a Shuttle mission. The KBB, as a component program of the MCC, serves to assist in the verification of the rule bases of the various expert systems, and to augment the training of the flight controllers. When complete, the KBB will verify and browse the CLIPS rule bases. This system, which in the view of its creators is a hypermedia system, will include the capabilities of automatic creation of links based on the CLIPS rule structure, querying the rules and saving the results as a collection, and browsing the rule bases either sequentially or by using the links and collections [POCK90].

The Space Station Freedom User Interface Language (SSF UIL). SSF IUL is in development at the Space Operations and Information Systems Division of the Laboratory for Atmospheric and Space Physics, University of Colorado, Boulder. It is designed for use by the astronauts, ground controllers, scientific investigators, and hardware/software engineers who will test and operate the systems and payloads aboard the space station. The IUL is object-oriented, English-like, supplements the graphical user interface to systems and payloads by providing command line entry, and will be used to write test and operations procedures. Hypertext is used to provide links between users of code (statements, steps, procedures, *etc.*) and associated annotation and documentation, linking code to object information, and linking steps within a procedure [DAVI90].

Artificially Intelligent Graphical Entity Relation Modeler (AiGerm). AiGerm is a relational database query and programming language frontend for Germ (Graphical Entity Relational Modeling) system. These systems are being developed by Microelectronics and Compute Technology Corporation, Software Technology Program (MCC/STP), Austin, Texas. There are three versions of AiGerm in use: Quintus Prolog, BIMprolog, and MCC's Logical Data Language (LDL). AiGerm is intended as an add-on component of the Germ system to be used for navigating very large networks of information, harnessing Prolog or LDL's relational database query capabilities. It can also function as an expert system shell for prototyping knowledge-based systems. AiGerm provides an interface between the programming language and Germ. When a user starts up AiGerm, the system builds a knowledge base of currently loaded Germ folio. The knowledge base is a collection of node, link, and aggregate facts. The user queries the database and runs programs that select, create, delete, inspect, and aggregate the nodes and links appearing in the Germ browser. To use AiGerm, the user first starts up Germ and loads the desired hypertext network folio into the Germ browser. In a knowledge base, for example, for each hypertext entity - *i.e.*, node, link, and aggregate - AiGerm asserts a fact (a prolog clause). AiGerm is currently used in MCC/STP's DESIRE (DESign Information REcovery) system to extract information on the design code for software systems. Research staff are experimenting with AiGerm in building IBIS (Issue Based Information Systems) -reasoning and decision support systems for software design and engineering. Rockwell International, an MCC/STP shareholder, uses AiGerm in a simultaneous engineering project. MCC/STP states that users of AiGerm can navigate Germ Networks or develop prototypes of knowledge-based hypermedia systems [HASH90].

PROJECT EMPEROR-I. This is a well known hypermedia project, merging microcomputer and videodisc hybrid technologies. It has been ongoing since 1984. It is a major research and development project which demonstrates how new technologies enhance better understanding and appreciation of a subject, in this case Chinese humanities, by delivering a large-scale online (real-time) hypermedia, multi-formatted, and multi-dimensional information simply not possible in sequential-formatted systems. The current hypermedia systems includes an interactive information delivery model for providing, at rapid speeds measured in fractions of of a second, requested relevent information in any format - visual, audio, textual

- as selected by the viewers at their pace and choice, including at the point of need. The project now includes:

o Two 12 inch NTSC CAV videodiscs, entitled "The First Emperor of China: Qin Shi Huang Di."

o Interactive courseware, at both a lay public and a serious researcher levels. Prototype courses have been developed for Digital Equipment Corporations' IVIS systems and for IBM PC compatibles. Later systems now include the Apple Macintosh Mac IIs.

o Electronic image databases for IBM compatibles and Mac IIs. Further development efforts have taken place with SOPHIATEC, Nice, France, and with the Project Athena of Massachusetts Institute of Technology involving a powerful multimedia image system using DEC's proprietary MUSE software for high-end machines such as DEC's MicroVax and IBM RTs. The EMPEROR-I hypermedia system has also been looked at for use on Sun3s and Sun4s.

o High resolution imaging digitization and electronic imaging has been performed on a Sun3-160 using OASIS software.

o Converting and creating large textual files with images and Chinese characters using MicroTek's MSF-3000 image scanner and INOVATIC's Readstar II Plus optical character recognition software. Digital textual files are kept in the hard disks, but when the data approaches 400-500 megabytes, CD-ROM can be produced.

This project, housed at Simmons College, Boston, Massachusetts, and aided by many interested resources, both industry and academic, is a masterful development. Professor Chin's goal is to show that computer power, storage technology, and software are now all available, at affordable cost, to provide the opportunities for innovative experimentation of ideas in education, training, research and development in nearly every subject field [CHEN88].

Intermedia. Brown University's Institute for Research in Information and Scholarship (IRIS), Providence, Rhode Island, has developed a powerful multi-user hypermedia software that allows professors, students, and other knowledge workers to create and follow links between electronic documents for different types. This system is named "Intermedia." This project defines hypermedia as the dynamic linking of data such that related data is easily accessible although the actual pieces of data may be stored in different physical locations. In theory the data can be any type, such as text, graphics, spreadsheets, video, or audio. Intermedia provides a desktop environment similar to that found on the Macintosh. The desktop contains applications (or tools) such as a word processor, a structured graphics editor, a historical timeline editor, a scanned-image view, an animation editor, a videodisk controller, and a viewer that displays and rotates three dimensional models. Users (now termed "viewer" or "authors"), with the tools just enumerated, enter data and link significant items of information together for a contextual viewing of that information.

Because of the extensive differences in the storage sources of the information, the Intermedia development incorporated two new concepts in the handling of the information, the "anchor" and the "proxy." The anchor concerns maintaining consistency across the applications; an anchor is a specific selection of data, a part of a document, with the surrounding information used to understand its significance. When a user follows a link, the document window opens to the size and location on the screen most recently saved, and automatically scrolls to the section that reveals the anchor with its surrounding information. The proxy is an intermediary concept used by the viewer for selecting an anchor in disparate data sources, *e.g.*, text, graphics, sound. The use of the data proxy concept allows the viewer to visualize non-graphical and conceptual media, to have simplicity in linking media, and to extend system applications to related data types [CATL88].

NASA STI Program Hypermedia Applications

This section analyzes the Clinical Practice Library of Medicine (CPLM) system developed under NASA Grant NAG10-0041 and partially funded by the NASA STI Program. The CPLM was conceived in 1979 by a team of medical and computer experts from the University of Florida and Kennedy Space Center. Since its onset, the system has evolved from a mainframe-based text database to a microcomputer-based hypermedia system that supports both text and high-resolution medical images. The design changes necessary to expand the system to include sound and animation are now being delineated.

The CPLM system is currently a computerized, rapid-reacting, medical reference system that could be placed aboard a long-term space flight to provide the spacecraft physician with nearly instantaneous access to the most complete medical references on Earth. With this type of support system, the physician could be confident that he was making the right diagnosis. The demonstration CPLM system that is available now runs on an IBM PS2 Model 80 microcomputer with a high resolution 8514A Display and a 1 gigabyte disk drive. The system is programmed in C under Microsoft Windows. The system contains a variety of medical texts including the STI Program's special publication NASA SP-3006, the "Bioastronautics Data Book." The CPLM system is written to allow expendability to the full capacity of the available storage device. Both traditional and hypermedia access to the information is permitted. Traditional Boolean search methods are enhanced by a parsings dictionary unique to each book that holds current spellings and root word divisions along with a lexicon that provides a book specific list of synonyms and abbreviations that automatically provides alternate search terms to the user. Word and phrase linkage among all documents is provided initially by the University of Florida project team with annotations to be eventually added by the physician end users. The educational capability of the CPLM system may be one of its major benefits in addition to its ability to deliver complex information in a user-friendly fashion.

Dr. Ralph Grams, University of Florida development team leader, stresses that the planned addition of voice activation, animation, and interactive hardware can make the CPLM system function as a fully automated physician's assistant. In a few years, Grams sees a miniaturized hypermedia CPLM system built into space suits and carried by Earthly physicians in their black bags. [GRAM91]

NASA STI Program (STIP) Hypermedia Plans

The NASA STI Program has put together a project plan to handle the development of a STIP Multimedia Initiative. The Multimedia Initiative plan covers both hypermedia applications fully controlled by the user and marketing applications that present multiple forms of information linked to each other with limited user control. The authoring system platform to be procured is a Macintosh IIfx equipped with 750 megabytes of storage, a videocassette recorder, a CD-ROM drive, a 35mm slide scanner, and hardware and software to support graphics, animation, sound, and real-time video capture and display. One of the most significant hypermedia applications planned within the STI Program is the NASA STIP Demonstration Electronic Performance Support System. This system will be developed at NASA STIP and will provide the proof-of-concept necessary to demonstrate how a performance support system can transcend the involvement spans of individual personnel integral to STIP projects. This demonstration hypermedia-based STIP associative memory will be used to illustrate the performance support concept to NASA scientists and engineers. Later phases of the performance support system will provide on-demand training in addition to a project knowledge base. In revamping its services to include multimedia, the NASA STIP will also develop procedures to handle information processing requirements such as cataloging of a new media and appropriate report documentation pages. In making a commitment to multimedia and hypermedia development, the STI Program acknowledges the application of hypermedia technologies that has begun throughout the STI community, and plans to foster the application of this technology to the removal of barriers to the transfer of scientific and technical information.

ACKNOWLEDGEMENTS

The authors wish to thank the many people who have assisted in their explorations of "Hyperprograms" to be considered for inclusion in the NASA Scientific and Technical Information Program (STIP). We also thank the Director of the STI Program, Ms. Gladys A. Cotter, for her direct support of our exploration, which has allowed us to look at many interesting areas in Hypermedia/Multimedia activities.

BIBLIOGRAPHY AND REFERENCES

[AIAA/NASA90] *AIAA/NASA Second International Symposium on Space Information Systems Proceedings, September 17-19 1990, Pasadena, California.* NASA/Jet Propulsion Laboratory, California Institute of Technology.

[BELL87] Christopher J. Bell. "A Review of Hypertext in a NASA Project Management Context." 1987. Johnson Space Center.

[BIEB90] Michael Bieber. "Automating Hypertext for Decision Support." *SEPEC Proceedings*, 1990 (See [SEPEC90]).

[CATL88] T. J. O. Catlin, and K. E. Smith. "Anchors for Shifting Tides: Designing a 'Seaworthy' Hypermedia System." *Online Information 88, Proceedings*, 1988 (See [ONLINE88]).

[CHEN88} Ching-chih Chen. "Hypermedia Information Delivery: The Experience of PROJECT EMPEROR-I." *Online Information 88, Proceedings*, 1988 (See [ONLINE88]).

[CHEN90] Ching-chih Chen. "Online Hypermedia Information Delivery." *National Online Meeting Proceedings - 1990* (See [ONLINE90]).

[CHRI90] Linda A Christman, Nam V. Hoang, and David R. Proctor. "Life Sciences On-Line: A study in Hypermedia Application." *SEPEC Conference Proceedings*, 1990 (See [SEPEC90]).

[CROU89] Donald B. Crouch, Carolyn J. Crouch, and Glenn Adreas. "The Use of Cluster Hierarchies in Hypertext Information Retrieval." *Hypertext'89 Proceedings* (See [HYPE89]).

[DAVI90] Randal L. Davis. "Hypertext As a Model for the Representation of Computer Programs." *SEPEC Conference Proceedings*, 1990 (See [SEPEC90]).

[GRAM91] Ralph R. Grams. *Final Report for NAG10-0041. 1990-1991, Clinical Practice Library of Medicine - Librarian and Tutor. Volume I.* University of Florida, College of Medicine, Department of Pathology, Medical Systems Group, Gainesville, Florida, 1991

[HAND90] T. Handley, Jr., and R. Masline. "High Rate Science Data Handling on Space Station Freedom." *AIAA/NASA Second International Symposium on Space Information Systems Proceedings* (See [AIAA/NASA90]).

[HASH90] Safaa H. Hashim. "AlGerm: A Logic Programming Front End for Germ." *SEPEC Conference Proceedings*, 1990 (See [SEPEC90]).

[HYPE89] *Hypertext'89 Proceedings, November 5-8, 1989, Pittsburgh, Pennsylvania.* The Association for Computing Machinery, New York, New York.

[LEVI91] Roger E. Levien. "A Civilizing Currency: Documents and Their Revolutionary Technology." *Technology 2000*, Massachusetts Institute of Technology Press, Cambridge, 1991.

[MARS91] R. Marshall. "Manipulating Full-text Scientific Databases: A logic-based Semantico-pragmatic Approach." *The Computer Journal*, Volume 34, Number 3, 1991.

[MOOR90] Jane Moorhead and Henry Brans. "Experiment Document Information System (EDIS) Evolution." *SEPEC Conference Proceedings*, 1990 (See [SEPEC90]).

[NIEL90] Jakob Nielson. *Hypertext and Hypermedia 1990.* Academic Press, Inc., San Diego, California, 1990.

[ONLINE88] *Online Information 88, 12 International Online Meeting, London 6-8 December 1988.* Learned Information (Europe) Ltd., Oxford, England.

[ONLINE90] *National Online Meeting, Proceedings -1990, New York, May 1-3, 1990.* Learned Information, Inc., Medford, New Jersey.

[POCK90] Tony Pocklington and Lui Wang. "A Knowledge Base Browser Using Hypermedia." *SEPEC Conference Proceedings*, 1990 (See [SEPEC90]).

[SEPEC90] *SEPEC Conference Proceedings. Hypermedia & Information Reconstruction:*
Aerospace Applications & Research Directions, December 3-5, 1990, Houston, Texas. University of Houston-Clear Lake, NASA/Johnson Space Center, Hypermedia Working Group.

[WARR90] Bruce A. Warren. "The Object the Metaphor the Power and Evergreen or The Eighth Way to Make a Hypermedia Project Fail." Warren-Forthought, Inc., [1991].

Automated Input into Databases: OCR and Descriptive Cataloguing

Gerhard E. Knorz
Fachhochschule Darmstadt
Department of Computer Science (IuD)
Schöfferstr. 3
D-6100 Darmstadt, FRG

Abstract

The paper deals with the technology of automated input from written text into databases.

The current emphasis in office automation and desktop publishing helped optical character recognition to become a useful and affordable method. Different approaches (pixel-oriented, feature-based, using dictionaries, using special algorithms for special problems[1]*, character transition probabilities or grammars) form the basis for meeting different classes of requirements such as demanded by small fonts, large varieties of fonts, special characters, different layout structures with mixed text and graphics, difficult printed matters with ligatures and kerning, poor printing quality.*

On the way to an automated (or semi-automated) input process from printed matters into databases[2] *OCR is nothing but a necessary step within the line of scanning, character recognition, descriptive cataloguing and (optional) content analysis. Cataloguing here is to be understood in a broader sense covering the identification and classification of the relevant pieces of input and the normalisation process according to database-specific rules.*

The role of efficient input into bibliographic databases is discussed, as well as problems and techniques of optical character recognition. AUTOCAT, a software prototype for automated cataloguing is introduced: its foundation, approach and user interface.

The project AUTOmatic CATaloguing (1985-1987) was sponsored by the BMFT under the contract no. 10200170

1 Information Management

1.1 Current Trends

The way to produce, organize and distribute information has been rapidly changing since the personal computer became a standard office tool. Word processing for text generation, business graphics and desktop publishing for information presentation, data bases for handling factual data and information retrieval systems for storing texts are in widespread use. Currently a major trend to computer-networks copes with the problem of communication and data exchange which nevertheless still is a problem. National and international standards like Office Document Architecture/ Office Document Interchange Format[3] (ODA/ODIF), Standard Generalized Markup Language[4] (SGML) and Electronic Document Interchange for Administration Commerce and Transport (EDIFACT[5]) are major efforts towards unhampered information exchange.

At present all the bright perspectives do not prevent paper from being the most important media of information interchange. A technology for transforming written text on paper into machine readable form is available and more and more being applied the last years. Even DER SPIEGEL, the Germans' most popular political magazine made a story out of this new trend, not without emphasizing remaining problems[6]. Information Retrieval Systems tend to offer an integration of an OCR-Software[7] package[8]. Effective applications in the field of office automation[9] and media documentation centers[10] demonstrate the feasibility and benefits of integrating OCR and text documentation.

[1] Broken or pasted up characters, underlined words

[2] Data interchange standards like SGML or ODIF form the alternative or supplementing strategy

[3] see [8]

[4] see [9]

[5] EDIFACT is going to become an international standard. In Germany, a draft is available as DIN 16556

[6] "Volles Rohr. Sogenannte Scanner, mit denen Gedrucktes direkt von der Vorlage in den Computer eingelesen wird, sind nun auch für den PC-Anwender erschwinglich." Der Spiegel 11/911, pp. 240 – 243

[7] Optical Character Recognition

[8] To give some examples: AskSam and Optopus, MegaStore and iBS GigaRead, Darwin and ReadStar

[9] For example at the Bundesanstalt für Flugsicherung in Frankfurt/M

[10] For example: at Gruner+Jahr, updating a database of journal articles or at BMW AG München, building a full text information system on new technologies(see [15])

1.2 The Importance of the Input into Bibliographic Data Bases

The user of bibliographic databases might not be aware of the fact that in some respects not the retrieval process but the input process actually is the critical operation[11]:

- not only economically, since at least more than 70 % of all costs of a documentation installation accrue from here,
- not only quantitatively, since a database like *PHYS* at the *Fachinformationszentrum Karlsruhe* has an input of about 10,000 new documents per month,
- but above all qualitatively, since topicality, completeness and reliability of a database as well as quality and stability of subject indexing are decisive criteria for a potential client.

If the costs of a database are high compared to the income, if these costs mainly originate from the input and if a lower level of the database' quality and quantity make the value of a database questionable, then powerful means have to be found to increase the effectiveness in this area. Clearly successful results have been achieved by co-operation and data exchange with producers of primary information and with other information offerers, by using profitable market prices e.g. for the service of writing offices and, last but not least, by a efficient organisation and by appropriate software with powerful functions (see [14]). In this paper, however, we shall focus on the argument that even better results can be achieved more effectively on the basis of OCR and expert system technology.

1.3 From paper to database records: a multistage-transition

Although hardware and software for an automated input into databases are available in principle, unexperienced users seem to underestimate this process. The DER-SPIEGEL-report, for example, describes some of the typical problems and it ends with the statement

> *"...even more effort than for the scanning process itself is necesarry to prepare the texts (12,000 pages) for a database[12]".*

A stepwise transformation process from written material into a machine readable database format, well prepared for retrieval, can be described as following (see fig 1):

Graphical level. A scanner is an input device, mapping the printed paper into a matrix of pixel elements (*pixel*). This matrix is stored as a file which could be manipulated by operations like those, offered by programs for drawing. The data structure supports its presentation as a facsimile representation of the original document at a high resolution screen. It cannot be used for retrieval without additional data.

Basicly a scanner can be specified by its resolution (dpi = dots per inch, i.e. pixels/area) and the data type of pixels: boolean (black/white), a spectrum of grey intensities or of colours.
Other features concern the size and the transport of the paper, and its ability to cope with different surface qualities of papers.

Character level. An optical character recognition software identifies collections of pixels to represent single characters (see section 2).

The result will be stored as an ASCII-file (or can be transformed into formats used by word processing or calculation software). It could be used for (sequential) searching, like editors do. But it is unable to differentiate between *Information Management* as a phrase in the title and the journal *Information Management* which might be the source of the document.

Formal document structure level. In general, bibliographic databases are based on a category scheme for documents. Relational databases use even more elaborate schemata. *Descriptive cataloguing* maps the input data into a representation fitting the predefined scheme. It is a well understood task within the framework of professional input of literature. It is to be understood in an even broader sense covering the identification and classification of the relevant pieces of input (for example to identify "Gerhard E. Knorz" to be the author of the paper) and the normalisation process according to database-specific rules (such as transforming "Gerhard E. Knorz" to "Knorz, G.E.").

Content level. Until now a document in a database consists of a set of categories each of which is composed by a sequence of words. Every single word is nothing but a character string between delimiters. Basicly, the standard technique of searching combines elementary search patterns by operators of Boolean Logic.

Manual content analyses aims to support the search for relevant documents by adding de-

[11] This section is mainly taken from [12], also dealing with automated input into bibliographic data bases but concentrating on aspects of automatic (subject) indexing.

[12] "Noch weit mehr Aufwand als das Scannen, so lernte der Strafverteidiger Kirch, würde es kosten, die Texte des Aktenkonvoluts für eine Datenbank aufzuarbeiten. Wenn der PC jedesmal die ganzen 12.000 Seiten hatte durchsuchen müssen, wäre das kein großer Fortschritt gewesen."

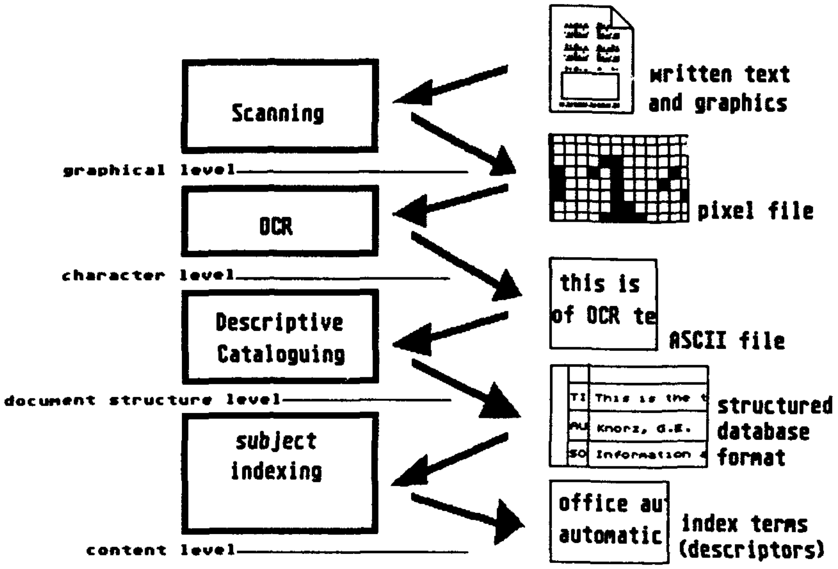


Figure 1: Four steps from written material into a data base

scriptors (often taken from a controlled vocabulary like a thesaurus) or classification codes. Automated indexing may simulate this approach (see [6]) or may support retrieval by means of other statistical or linguistical techniques.

This decomposition of an automated input process into single steps reflects the underlying conceptual structure of the task. Depending on the degree of software integration, the user will not realize the first or even the second step to form a independent step. Normally, the third step is very poorly performed: the results of the OCR-processing are assumed to be the content of one single data field (category) and the document record is manually supplemented with additional data. Section 3 describes an experimental but well elaborated solution. In [11] p. 59 a new software product is announced, which seems to realize a very simular approach.

For specialized and restricted applications dedicated solutions make sense: i.e. the OCR-Software *Gigaread*, especially developed for reading adress books, identifies the different parts of an adress item and forms a structured database entry as a result.

The fourth step can be discussed quite independently from the aspect of applying OCR-techniques. We do not look into this aspect here (further readings: [12], [13] or [6])

2 OCR — problems and techniques

2.1 Types and sources of potential problems

Some typical problems will be summarized[13] in short. Documents often make use of a variety of fonts and sizes. Expecially small (6pt and smaller) and large fonts (24pt and larger) may cause problems. The same applies to special fonts[14] (i.e. italics), special characters, a large number of different fonts within one document, very small distances between characters, kerning, the use of ligatures and of underlining.

Other types of problems arise from (poor) printing quality: characters may be broken into pieces, causing "m" to look like "rn" for example. On the other hand two characters may merge into one. Small pieces of characters may be omitted (like the upper part on the "i" or may be present as some kind of noise only. Handwritten additives may interfer with original lines of text as well as background noise and patterns.

The more creative the layout, the more difficult is the correct interpretation. We can expect graphics and text to be separated automatically (or manually

[13] The program committee recommended to concentrate on the aspect of descriptive cataloguing. Accordingly other topics had to be shortened.

[14] A particular type of problem with some fonts is the destinction between "1", "l" and "I"

controlled). Columns of text are to be recognized and read in a meaningful sequence.

For further processing the information on the original fonts and sizes may be preserved or not, depending on the system used and the needs of the application.

2.2 Approaches and techniques to character recognition

A variety of stategies try to cope with the requirements of different application types.

Pattern matching algorithms basicly compare the pixel image of a character with standard images out of a training set of characters. These algorithms have the advantage to be of a fast performance and to be trainable to arbitrary fonts and character sets. The disadvantage is that they are not independent of character size and font, so normally a training phase must procede the reading phase.

Feature recognition based algorithms try to identify the interpretation of a character by means of detected fearures[15] (loops, crossings, lines). They depend on fixed character sets, but are quite independant from size and fonts. Normally, training is not necessary and sometimes not even possible.

Training of unknown characters or known characters with unfamiliar shapes may be necessary, possible or impossible. A dictionary may be used for (additional) automated training[16].

A conflict set of characters, as a special kind of training set, might support the discrimination of similar characters.

Context analysis can be applied to introduce arguments of the type: the letter "O" in the context of numbers most likely turns out to be a "0". Another method is to apply a pattern matching filter a posteriori, detecting typical recognition errors in the produced text file.

The system AUTOCAT, described in the next section, made use of the OCR workstation KDEM 1.200. This machine can be trained for different fonts and sizes and it preserves this kind of information together with the recognized character itself for further processing. KDEM 1.200 successfully has been tested in the course of the AUTOCAT project and it was the only appropriate OCR system available at that time (1985). In this aspect, the situation changed totally.

3 An Expert system approach to automated cataloguing

The result of formal cataloguing is a formally structured document description, derived from the primary publication[17]. In order to perform the tasks in the fields of library and documentation and, last but not least, to allow data exchange between different information facilities, extensive sets of rules like the Anglo-American Cataloguing Rules (AACR 2,620 pages) for example were established. The present high degree of explicit standardization and the fact that this is a non-trivial labour-intensive task including a great share of routine situations suggest the development of a knowledge-based system for formal cataloguing. This idea was taken up at different places in the last decade (see [3]). Simpler systems only support the standardization of data entries. If a knowledge-based system is supposed to cover the whole process of formal cataloguing [10], [17], [18], then the interest must focus on the first step, namely the identification and categorization of different elements of the input.

AUTOCAT[18] (AUTOmated CATaloguing), to be described in more detail, means both, project and system (see [?]). It certainly was one of the largest efforts in the field and made a serious attempt at practicle applications.

- It adheres to the relatively simple cataloguing rules of the International Nuclear Information System *INIS*.
- It relies on knowledge about document types, in particular on an empirical information structure of physics journals.
- The prototyp of a cataloguer's working place has been developed providing a user interface that allows to control the system and to correct its output.

AUTOCAT was developed in two phases: The first phase at the Technical University of Darmstadt ended with a Prolog prototype that showed the feasibility of the relatively simple cataloguing of articles from physics journals. In its second phase, a software house took over the project. During this time AUTOCAT made some steps towards a marketable software product. This prototype was presented at the Hannover Fair '89.

[15] Some produces try to establish the phrase *Intelligent Character Recognition (ICR)*

[16] The characters of recognised words, found in the dictionary, are chooeen as a training set. So, the systems adopts the specific features of the copy on hand. It increases reliabiltity and speed.

[17] A much more detailled publication on this section's subject is going to be completed ([5]). The following material is adopted from this publication.

[18] The project AUTOmatic CATaloguing (1985-1987) was sponsored by the BMFT under the contract no. 10200170

3.1 The AUTOCAT approach

AUTOCAT produces records for a bibliographic database and its prototypical application environment is INIS. In contrast to cataloguing assistants, AUTOCAT comprises the whole process of descriptive cataloguing.

AUTOCAT was first developed for cataloguing articles from physics journals. *Altough its concept* has been extended to report and monograph cataloguing, the AUTOCAT approach still is best explained using the core application for scientific journals.

AUTOCAT catalogues periodical articles in two main steps:

1. It recognizes information elements in the machine readable jcurnal as detailled as it is necessary for cataloguing under INIS cataloguing rules.
2. It normalizes the information elements as stipulated by INIS rules and enters them into the target record, defined by the categories of the INIS worksheet.

AUTOCAT starts its work with a first representation of the journal article, generated as output of an OCR-processing step. Simple layout features like fonts and distances of tokens are used to detect input blocks that are candidates for functinal roles *like title, author* etc. (see fig. 2). The targets of recognition are defined by the information structure of the journal under processing. Network grammars are bound to that target structure (see fig. 3) and interpret the *input representation with the help of*

- the known information structure of articles in the journal augmented with the relevant items of the journal itself.
- lists of keywords for special categories like affiliations.

The categories of the INIS worksheet form the final target structure, to which normalization rules are bound. They capture their input from the stored intermediate results and transform it in adherence to INIS rules.

3.2 The AUTOCAT representation of physics journals

Since cataloguing rules like AACR2 tell a cataloguer what to look for, they must define abstract expectations about the journal or any other document type: So, implicitly they define a normalized informational model of the document type. On the basis of such an abstract document model, one can characterize and recognize an occuring document as a realization of an abstract information object. The AUTOCAT representation of journals' information structures expands these ideas. It uses categories taken from cataloguing rules like INIS or AACR2 (e.g. title proper, author, affiliation), completed only as far as necessary by other constructs that cater for observed data (e.g. head of article or page frame). The representation, developed by AUTOCAT, is the result of an empirical investigation of 330 core journals of physics. In a recent copy of every journal, title page, table of contents and the first and the last page of at least one paper were exploited for data about the occuring information elements, their sequence, structure cues like separators, fonts and spaces. An additional test using complete volumes of 40 journal's complete volumes supplied reassuring results: the information structure of physics journals is reasonably stable. (see [2]).

Some simple observations help to specify the technical features of a representation device for information objects as it is needed here:

Since information objects are mainly defined by their composition, the **part-whole relation** will be the backbone of the representation. Generic relations appear less important.

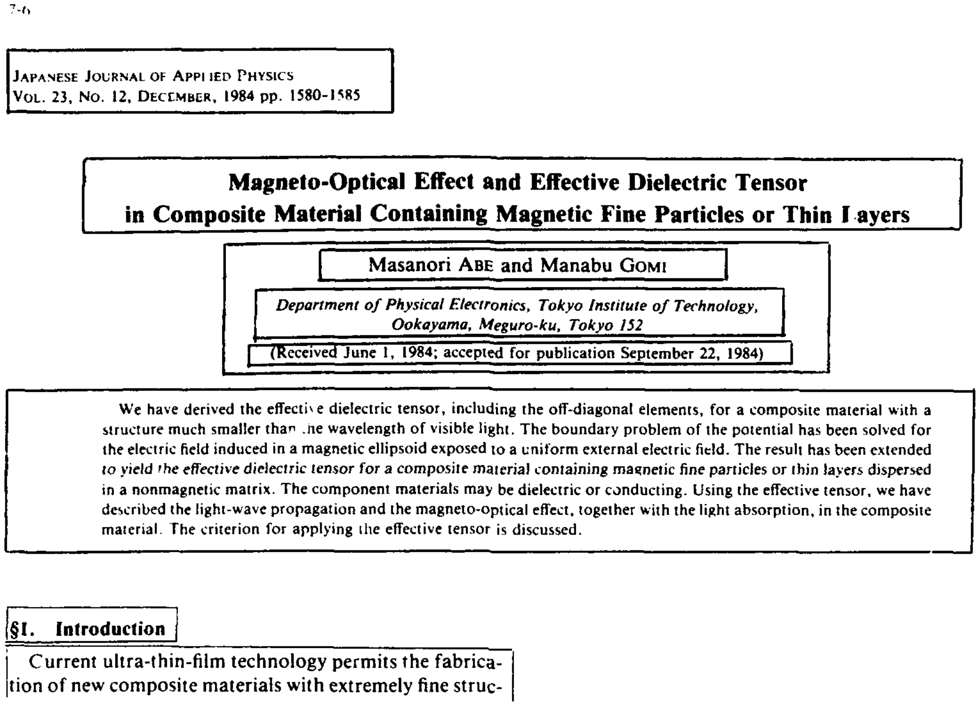
"Natural" objects in a reader's eyes, like an article or a table of contents, may be very large in terms of knowledge representation constructs (such as frames). A facility for the adequate representation of **"complex objects"** must be provided.

An information object may have multiple instances: A paper may have, for instance, a couple of authors, but not a couple of titles. So, the description of individual elements has to include a **cardinality restriction**.

Information objects may occur repeatedly for the reader's orientation, and thus provide useful redundancy for the recognition process. It must be possible to state **equivalence relations** of this type between information objects.

As recognition means to find an instance of an information object in the actual document copy, it is reasonable to associate **recognition methods** to appropriate information objects.

Scientific journals differ in their presentation without being heterogeneous: They combine the available design solutions in an individual way. The necessary individualiziation of periodicals can be achieved by **typing their components**. The individual journal is described by the set of types of its parts. It must be possible to store this type of information and include it into the abstact representation of the information structure before processing.



7-6

JAPANESE JOURNAL OF APPLIED PHYSICS
VOL. 23, NO. 12, DECEMBER, 1984 pp. 1580-1585

Magneto-Optical Effect and Effective Dielectric Tensor in Composite Material Containing Magnetic Fine Particles or Thin Layers

Masanori ABE and Manabu GOMI

Department of Physical Electronics, Tokyo Institute of Technology, Ookayama, Meguro-ku, Tokyo 152

(Received June 1, 1984; accepted for publication September 22, 1984)

We have derived the effective dielectric tensor, including the off-diagonal elements, for a composite material with a structure much smaller than the wavelength of visible light. The boundary problem of the potential has been solved for the electric field induced in a magnetic ellipsoid exposed to a uniform external electric field. The result has been extended to yield the effective dielectric tensor for a composite material containing magnetic fine particles or thin layers dispersed in a nonmagnetic matrix. The component materials may be dielectric or conducting. Using the effective tensor, we have described the light-wave propagation and the magneto-optical effect, together with the light absorption, in the composite material. The criterion for applying the effective tensor is discussed.

§1. Introduction

Current ultra-thin-film technology permits the fabrication of new composite materials with extremely fine struc-

Figure 2: Header of an article with marked blocks

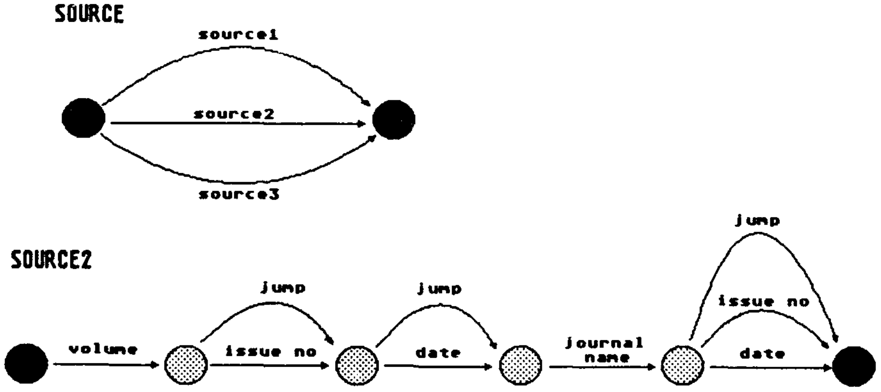


Figure 3: A simplified Network Grammar(Discussion in the subsection on grammars)

For the implementation of a representation that copes with these requirements, a frame language can be used, embedded into a knowledge engineering environment.

3.3 Selected aspects of the AUTOCAT prototype

3.3.1 Grammars for analyzing information elements

AUTOCAT used hierachies of ATN-like transition grammars (see fig. 3). The nodes and links are written in a user-oriented Prolog notion which subsequently is translated into a Prolog module. A grammar takes two types of parameters: a specification of the input (e.g. an author name grammar must not cross a block broader[19]) and a selection of a particular alternative (sub-grammar for a particular object type). The results are transferred via special output registers.

We can make a difference between types of structures for which grammars are to be defined. It is necessary,

- to specify the structure of document types, e.g. page layout, block structure (**macro structure**)
- to specify the syntax of different information objects, e.g. author block, affiliation block (**micro structure**)

[16] describes the design of an GEM-based[20] implementation of a graphical interface (see fig. 4), which can be used to compile an actual macro structure of a document type from a set of single components. The user models a document type as a sequence of pages. (S)he selects primitive or composed information items like title, abstract, author and so on and places them on the page. Each object (page, title, etc.) is represented by a named box. This way, the presentation of the part-whole-relation is quite natural. The user is able to formulate alternatives and may define a hierarchy of substructures. The final schema representation of a real document page is translated into Prolog facts.

From our experience with grammars we do not feel that they form a reliable tool for being used by the enduser: Complex grammars as are necessary for analyzing the micro structures of this application cannot be modified without extensive testing. It is not possible to decide whether a rule is absolete, wrong, or correct, just by looking at a simple rule which seems to cause problems, without recognizing the context of the whole grammar. In [7] another way for defining syntactical structures is proposed and implemented. It starts from the observation that a user might modify a grammar typically in a situation where the system cannot cope with a particular example. Why should this very example not be used to teach the system? So the grammar as the internal representation of syntactical structures is incrementally constructed from a set of examples. The user types in the example, the system breaks down the input to a sequence of tokens.

The user has to decide for which token in the example which level of abstraction is appropriate. Afterwards (s)he has to teach the system the constituent structure of the examples (see fig. 5). The system has to transform the information given by the user into grammar rules and has to ensure that no conflicts with existing rules arise. The implementated version of the system translates each example into one corresponding (complex) grammar rule. This reference to an actual example should tell the user much more on the reason of an underlying problem than an abstract grammar rule.

3.3.2 The user interface of the Knowledge Craft prototype

The shift from journal cataloguing to report and monograph cataloguing caused some substantial changes in the architecture of the AUTOCAT system. A basically frame-oriented expert system development tool, Knowledge Craft, is used for the implementation. Knowledge Craft offers backwards chaining (CRL- PROLOG) and forward chaining rules (CRL-OPS). In the following, we shall focus on the user interface.

The interface handles the interaction betweeen the user (cataloguer) and the AUTOCAT system. There are several functions which should be provided by the user interface. The main task is to control and correct the cataloguing results. To control the results, a document facsimile is reconstructed on the basis of the stored layout information, and a representation of the corresponding INIS- worksheet entries is displayed on the screen. To correct the results, the user must be able to edit the worksheet entries.

The initial window on the screen contains a list of all newly processed documents which have not been handled by the cataloguer. The documents are classified into recognized and unrecognized documents. *'Unrecognized'* is a label for documents with obligatory worksheet entries missing[21]. Clicking at a document identification with the mouse, AUTOCAT selects the document. For interaction, several command buttons are offered (e.g. edit-worksheet, change-level, next-page).

[19] Blocks and its broaders are defined in a preprocessing phase by means of layout features.

[20] Graphics Environment Manager. A trademark of Digital Research Inc.

[21] It is possible that AUTOCAT was not able to find all the necessary information in the input.

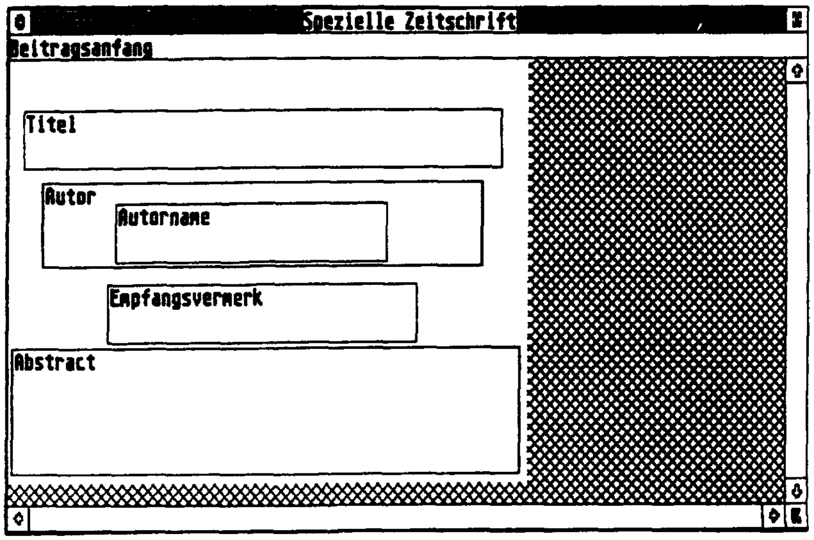
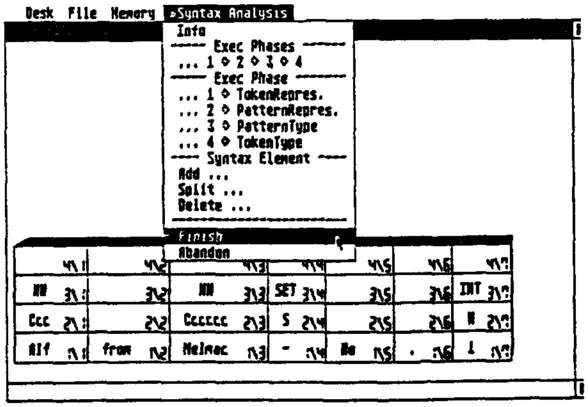


Figure 4: A simple makro structure definition, designed on a graphical interface



4\1	4\2	4\3	4\4	4\5	4\6	4\7
NN 3\1	3\2	NN 3\3	SET 3\4	3\5	3\6	INT 3\7
Ccc 2\1	2\2	Cccccc 2\3	S 2\4	2\5	2\6	N 2\7
Alf 1\1	from 1\2	Melmac 1\3	- 1\4	No 1\5	. 1\6	1 1\7

Figure 5: "Alf from Melmac – No. 1" as an example of the form (noun) (from) (noun)(special character) (No) (.) (integer). The interpretation is *name (al(who(Alf),loc(Melmac)),gen(no(1)))*.

The user can edit the worksheet or invoke an intelligent copy function by activating a worksheet tag and subsequently marking a passage of the document faximile[22].

The selected passage is transformed according to INIS rules and copied into the active entry.

4 Conclusions

Both, the Prolog and the Knowledge Craft prototype contribute to different aspects on the way to an automatic cataloguing system suitable for practical application. It took about 3 years for a univerity research group to develop the foundations and the first Prolog prototype. It showed the feasibility of the relatively simple cataloguing of articles from physics journals on a broad empirical basis.

It took about half a year, to designe and implement the Knowledge Craft prototype. Its purpose was twofold: It could be demonstrated, that knowledge-based automatic cataloguing can serve as a helpful cataloguers' tool if it provides an appropriate user interface. Furthermore, the implementation showed exploratorily that the modified approach for recognition supports more difficult cataloguing tasks.

More generally speaking, the AUTOCAT project demonstrated the soundness of knowledge-based concepts in the field of the information market. It showed that OCR and expert system technology may very well help to solve the practical problem of the information industry.

The potential applications of the AUTOCAT technology is not limited to the formal cataloguing of document types described in this article. The techniques developed during the AUTOCAT project are also applicable for other document types, e.g. telex, electronic mail, technical documentation.

Quite naturally, still there are limitations. Whether or not current OCR technology can meet specific application demands has to be proven by carefull pretests. But they stand a good chance. On the other side, AUTOCAT's document type is quite static (compared with the flexibility supported by SGML and ODA). Nevertheless, it should be clear that the demands of routine tasks can be fulfilled on the basis of available methods and technologies.

Acknowledgements

This contribution is based upon the results of different projects sponsored by the BMFT (German Department of Research and Technology), in which many colleagues were engaged.

As indicated, the section on AUTOCAT is a shortened and modified version of parts of a paper which I am currently preparing together with several coauthors: Brigitte Endres Niggemeyer, Ulrike Rauth and Heinz Marburger.

I want to express my sincere thanks to all persons involved.

References

[1] Appelt, W., "Normen im Bereich der Dokumentverarbeitung", Informatik-Spektrum (1989)12: pp. 321–30

[2] Below, A. v., "Funktionsgerechte Strukturierung von Fachzeitschriften", Nachr. Dok. 38(1987), pp. 343–349

[3] Davies, R. and James, B., "Towards an Expert System for Cataloguing: some Experiments Based on AACR 2", Program 18 (1984) 4: pp. 283-297

[4] Endres-Niggemeyer, B. and Knorz, G.: "AUTOCAT: Wissensbasierte Formalkatalogisierung von Fachzeitschriften", in: Brauer, W.; Wahlster, W. (eds.): "Wissensbasierte Systeme", 2nd Int. GI-Kongress 1987 Berlin: Springer, 1987, pp. 53 – 62.

[5] Endres Niggemeyer, B.; Knorz, G.; Marburger H.; Rauth, U., "AUTOCAT — a Knowledge-Based Approach for Descriptive Cataloguing". To appear.

[6] Fuhr, N,; Hartmann, S.; Lustig, G.; Schwantner, M.; Tzeras, K.; Knorz, G., "AIR/X — a Rule-Based Multistage Indexing System for Large Subject Fields", in: Research and Development in Information Retrieval (Proc.),1991

[7] Götz, P., " 'Learning by Examples' als Strategie zur Wissensaquistion. Entwurf und Implementierung einer Benutzeroberfläche zur Beschreibung syntaktischer Strukturen", Diploma Thesis. TH Darmstadt, Dep. of Computer Science, 1989

[8] "Information Processing – Office Document Architecture (ODA) and Interchange Format", ISO 8613, International Organization for Standardization, Geneva, 1989

[9] "Information Processing – Standard Generalized Marup Language (SGML)" ISO 8879, International Organization for Standardization, Geneva, 1989

[22] The inclusion of a snapshot of the interface had to canceled because of poor image quality

[10] Jeng, L.-H., "An Expert System for Determining Title Proper in Descriptive Cataloguing: A Conceptual Model", Cataloguing & Classification Quarterly, 7(1986)2: pp. 55 – 70

[11] KI. Künstliche Intelligenz: Forschung, Entwicklung, Erfahrungen Organ der Fachbereichs 1 der Gesellschaft für Informatik e.V. (GI). (1990)4

[12] Knorz, G., "Automatic Cataloguing and Indexing", In: Barnev, P.; Kerpedjiev, S (eds.), "Proc. of Programming '90", Sofia, Bulgaria, 1990.

[13] Lustig, G. (ed.), "Automatische Indexierung zwischen Forschung und Anwendung", Olms, Hildesheim, 1986.

[14] Marek, D., "Zwei Jahre Online-Input im Fachinformationszentrum Energie, Physik, Mathematik", ABI-Technik 3 (1983): pp.. 201-208

[15] "Volltext-Datenbank per Scanner erstellen", PC-Magazin, 17(1991), apr. 1991

[16] Pitz, H., "Beschreibung der Informationsstruktur von Zeitschriften für ein wissensbasiertes System zur Formalkatalogisierung. Entwurf und Implementierung einer Benutzeroberfläche", Diploma Thesis. TH Darmstadt, Dep. of Computer Science, 1987

[17] Schirra, J. R.; Brach, U.; Wahlster, W.; Woll, W., "WILIE - Ein wissensbasiertes Literaturerfassungssystem", in Endres-Niggemeyer, B.; Krause, J. (eds.): "Sprachverarbeitung in Information und Dokumentation", Berlin: Springer 1985, pp. 101–112

[18] Weibel, S.; Oskins, M.;Vizine-Goetz, D., "Automated Title Page Cataloguing: A Feasibility Study", Information Processing & Management, 25 (1989) 2: pp. 187-204

Data Compression Techniques

R.A. Hogendoorn
National Aerospace Laboratory
P.O. Box 153
8300 AD Emmeloord
The Netherlands

1 Summary

Data compression can be used to reduce the volume of documents. This results in considerable savings on storage capacity and in transmission time. In general, there are two classes of compression techniques: reversible compression or statistical compression, with which documents can exactly be reproduced, and non-reversible or noisy compression, with which documents can be reproduced up to a given fidelity. Non-reversible compression most often gives a higher compression factor, but, obviously, at a price. Reversible algorithms are suitable for text compression and black-and-white images. Non-reversible algorithms are suitable for images. This paper describes the advantages and caveats of data compression. What is to be expected if data compression is used and, more important, what is not to be expected.

2 Introduction

Data compression is gradually becoming accepted as a means to make better use of available resources like communication channels and disk storage. For example, facsimile would not quite have been as useful as it is now if data compression had not been used. Facsimile uses an algorithm called the modified READ (Relative Element Address Designate) code which compresses a digitally scanned page by a factor of 13 on the average. This means that the transmission time is only about 1 minute instead of 13 minutes.

Another example, in which data compression is indispensable, is High-Definition TeleVision (HDTV). The raw data rate coming from a HDTV camera is about 470 Mbit/s, whereas the television broadcast channel will only allow a data rate of about 1.75 Mbit/s. In order to obtain this large reduction of the data rate, quite elaborate data compression techniques have to be used.

Now, what is data compression? Essentially, data compression is a matter of modelling: one tries to model the mechanism that generated the data, i.e. the source. Hence, data compression is also called "source coding". Of course, the better the model is, the less information is needed to describe the data, i.e. the output generated by the source. Note that it is assumed that the model is known at all times. The fact that data compression is modelling explains why there are so many different data compression algorithms. Each algorithm has its own peculiarities, which makes it more, or less, suited for a particular application.

Data compression algorithms can be divided into two classes, namely reversible and non-reversible algorithms. Reversible algorithms only change the representation of the data into a more efficient one. The data, resulting from decompression, are identical to the original data as generated by the source. Non-reversible algorithms, on the other hand, make only an approximate representation of the original data. The result after decompression differs from the original data by a certain amount, called the distortion. For each non-reversible algorithm, there is a trade-off between the amount of distortion and the compression factor. A larger compression factor results in a larger distortion.

One of the fundamental questions of data compression is what is the shortest possible representation of the output of a source? To answer this question, the concept "information" needs to be defined more precisely. Basically, the amount of information that is conveyed by a symbol depends upon the predictability of the symbol, i.e. its probability of occurrence. The less probable the symbol, the more information is conveyed by it. Using this notion, the amount of information generated by a source can be defined:

$$H_0(x) \triangleq \sum_{i=0}^{N-1} -\Pr(x=i)\log\Pr(x=i) \qquad (1)$$

$H_0(x)$ is called the "entropy" of the source. The entropy is maximum for a source that outputs equiprobable symbols, i.e.

$$\Pr(x=i) = 1/N,\ i \in \{0,\ldots,N-1\}$$

An example of a source is the throwing of a die. The output symbol is the number of spots that shows up. For this source, the entropy

$$H_0(x) = 2.59 \text{ bits/symbol.}$$

If the die is loaded, say, the probability of throwing a 1 is 1/2, then

$$H_0(x) = 2.16 \text{ bits/symbol.}$$

In equation (1), it is assumed that subsequent output symbols of the source are independent of each other. Normally, this is not the case. For example, the entropy of the english language is estimated to be

$$H_0(x) \approx 4.5 \text{ bits/letter}$$

if letters are considered independent of each other. However, if the inter-letter dependencies, like the fact that a "q" is almost always followed by a "u", are taken into account, the entropy is estimated to be

$$H_\infty(x) \approx 0.6 \ldots 1.3 \text{ bits/letter.}$$

Up till now, only discrete sources, i.e. sources that can output symbols from a finite alphabet, have been considered. There are also sources that output real values, e.g. the reading of a current meter. In such a case, the accuracy of the reading has to be taken into account to define the equivalent of an entropy. This results in a curve, the so-called "rate-distortion curve" The rate-distortion curve specifies, for each distortion, the entropy of the source. Note, that no particular distortion measure is specified in advance; it depends entirely upon the application that makes use of the decompressed data.

Why is the entropy such an important concept? Well, a main result in source coding theory is C. Shannon's *Source Coding Theorem*. Essentially, this theorem states that it is not possible to find error-free codes with an average length below the entropy of the source. More specific, let sequences of L source letters be coded in N bits and let only one source sequence correspond to each code sequence. Let Pr(error) be the probability of occurrence of a source sequence for which there is no code sequence. Then, for any $\delta > 0$, if

$$N/L \geq H_L(x) + \delta \qquad (2)$$

Pr(error) can be made arbitrarily small by making L sufficiently large. Conversely, if

$$N/L \leq H_L(x) - \delta \qquad (3)$$

the Pr(error) must become arbitrarily close to 1 as L is made sufficiently large.

A similar result holds for source coding with a fidelity criterium. It is not possible to find codes with a rate-distortion curve that lies, at any point, below the rate-distortion curve of the source. Note that the rate-distortion curve not only depends on the source, but also on the distortion measure. This means that for the same source, several different rate-distortion curves are possible depending on the distortion measure that is used.

3 Reversible Data Compression

Reversible data compression algorithms change tne representation of the information such that the average length of the representation is as short as possible. This is done by using information about the probabilities of occurrence of source symbols. A simple example of such an algorithm is *run-length coding*. A run-length encoder replaces series of consecutive identical symbols by the symbol followed by a repeat count. Especially with black-and-white images, this simple algorithm gives a fair amount of compression.

A more elaborate coding technique is *Huffman coding*. A Huffman encoder uses a fixed table of codewords. The codewords are chosen such that the symbol with the highest probability of occurrence gets the shortest codeword, the symbol with the second highest probability gets the second shortest codeword and so on. Huffman codes are good codes: Let $\bar{l}_L(x)$ be the average length of the Huffman code that codes L source symbols at a time, then

$$H_L(x) \leq \bar{l}_L(x) \leq H_L(x) + 1/L \qquad (4)$$

In other words, Huffman codes result in an average codeword length that is arbitrarily close to the theoretical minimum by taking L large. However, a large L results in such a large codeword table that it is impossible to implement such a code. Also, it is assumed that the *probabilities of occurrence of symbols do not change*. In practice, this is not always true. Huffman codes cannot efficiently adapt to these changing probabilities. In a worst case situation, even data expansion may occur. For example, if the digital facsimile compression algorithm is used for halftone (grey-level) images, data expansion occurs, since the codeword tables were designed for binary images. Although Huffman coding has its limitations, it is an important coding technique that is often used as a part of a more complex compression algorithm.

A coding technique, that overcomes the limitations of Huffman coding is *arithmetic coding*. It is more complex than Huffman coding, but it is also more efficient and it can adapt to changing statistics quite easily. Arithmetic coding is used in the newer data compression algorithms, such as the JPEG (Joint Photographic Experts Group) standard algorithm for still images.

An entirely different approach is *dictionary coding*. The idea is to have a dictionary of strings (series of symbols) of possibly different lengths and, instead of the string, the index of the string in the dictionary is sent. The *Ziv-Lempel* algorithm is an example of this class of algorithms. The Ziv-Lempel algorithm builds its dictionary adaptively, based upon the past input symbols. There are many variations upon the basic Ziv-Lempel algorithm; they all give a good compression and can be easily decoded. Therefore, this type of algorithms is popular for compression of files on personal computers and workstations.

4 Non-Reversible Data Compression

Non-reversible data compression algorithms make only an approximate representation of the data. They introduce a distortion with respect to the original data that depends on the compression factor. The larger the compression factor, the larger the distortion. Non-reversible compression algorithms are also called "noisy" or "lossy" compression algorithms. The advantage of non-reversible compression algorithms is that with these algorithms much larger compression factors can be obtained than with non-reversible algorithms. Furthermore, there are also practical reasons for using non-reversible algorithms. Reversible algorithms assign codewords depending on the probabilities of occurrence of source symbol sequences. If N is the number of different source symbols and L the length of the sequence of source symbols that are encoded in one codeword then the size of the codeword table is N_L. Obviously, the codeword table size is far too large for anything but small values of N and, especially, L. Now, given that $N = 256$ for most images (8-bit pixels),

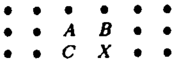


Figure 1: Image pixels used in prediction

it is clear that the coding of images is almost exclusively done with non-reversible algorithms.

The following paragraphs describe the most important non-reversible compression techniques assuming that the data being compressed is image data, i.e. a picture consisting of $L \times L$ pixels.

One of the most simple techniques is predictive coding. The idea is to use the values of neighbouring pixels to get an estimate of the value of the current pixel. Instead of using the current value of the pixel, the difference between the estimate and the current value is used. If the picture does not change a lot from pixel to pixel, the differences can be encoded far more efficiently than the original pixel values, e.g. by using a suitable Huffman code. Predictive coding is not inherently non-reversible, since the differences can be coded exactly. However, by disregarding all differences below a certain value, the compression factor can increase a lot without a significant loss in picture quality. Figure 1 shows the pixels used in a simple predictor. The estimate $\hat{X}$ is

$$\hat{X} = B + C - A$$

The picture is encoded row by row and only the very first pixel value is needed to start the decoding process, so this value is retained with the compressed differences. Predictive coding gives a fair compression, while the implementation complexity is low.

A more elaborate technique is *transform coding*. By using a mathematical technique, called "transform", the picture is split into components that describe the variations in the picture, i.e. there is a component that specifies the average pixel value up till a component that specifies the fastest variations, that is pixel-to-pixel variations. Transformation requires a fair amount of processing time. Therefore, it is done on small parts of the picture, e.g. blocks of 8×8 pixels. Also, special transforms are taken, like the Discrete Cosine Transform (DCT), because they are computationally more efficient than the optimum transform.

The reason to look at the variations in pixel values, rather than the pixel values itself is that these variations are most often very gradual. Therefore, a few transform components are sufficient to give an accurate reconstruction of the block of pixels. The distortion measure used with transform coding algorithms is the Root-Mean-Square (RMS) error. This error can easily be expressed in terms of the dropped transform components. So, it is quite easy to make an adaptive algorithm, with a guaranteed maximum RMS error. Transform coding gives a very good compression, but is expensive in terms of computations and more elaborate to implement. The JPEG standard algorithm for images is a transform coding algorithm.

Vector quantisation (VQ) is a compression technique that, in theory, can perform quite close to the rate-distortion bound. The image is divided into vectors, for example, a $n = l \times m$ block of pixel values can taken as a n-dimensional vector. Each vector is then compared with a collection of representative templates or *codevectors*. The best match is chosen and its index in the codebook becomes the codeword. Essentially, VQ is a dictionary method as described in the previous section.

Compression is obtained in VQ by using a codebook with relatively few codevectors compared to the number of possible vectors. The complexity of VQ depends on the size of the codebook and the way in which comparisons with the codevectors are made. Compression is very good and decoding is very fast. Therefore, it is often used for coding video images. A disadvantage is that the generation of a good codebook requires a lot of processing.

5 Applications

At this moment, data compression techniques for text compression are readily available on most workstations and personal computers. A large share of the algorithms are in the public domain and can freely be used. Most of them are based on the Ziv-Lempel algorithm and give a fair compression (factors between 2 to 5). Several of these algorithms have become de-facto standards, like /em compress and *lharc*. For other types of information, like images, speech and video, only "canned" and application-specific solutions are available. This situation, however, will change within the next few years. A lot of standardisation has been and is being done. JPEG specified a standard algorithm for still images. The Motion Picture Experts Group (MPEG) prepared a draft standard algorithm for video. It will only take a couple of years before these standard solutions are available on a wide basis and for an affordable price. The development of HDTV, will also result in more and better data compression techniques. All these techniques, however, will be variations upon the basic algorithms, described in this paper. The evolution of micro-electronics makes it possible to use the more advanced and computationally intensive compression algorithms that give far better compression than the present algorithms.

COMPUTERIZED PROPERTY DATA FOR ENGINEERING MATERIALS
An Overview

Gordon H. Wood
Manager, Canadian Scientific Numeric Database Service
Canada Institute for Scientific and Technical Information
National Research Council of Canada
Ottawa, Canada, K1A 0S2.

SUMMARY

The special problems of materials property databases, as opposed to scientific numeric databases, are described. A review of progress towards materials property databases since the seminal workshop of 1982 in Fairfield Glade, Tennessee is given based on the recent Third *International Symposium on the* Computerization and Use of Materials Property Data. Topics include standards and data representation, standards and database development, expert systems and materials databases, data issues for engineering materials, industrial applications, and working and prototype systems.

1. INTRODUCTION

In the course of daily life, few people give much thought to the engineering effort invested in the structures and devices which they use and trust. Flying at 10 000 metres in a modern aircraft, turning on a new electrical appliance, swinging a high-tech tennis racquet at a ball -- all these are acts of faith in an integrated design and manufacturing system whose output can be trusted to be safe and reliable. That, no doubt, is as it should be; those concerns are delegated to experts whose task it is to take both natural and man-made materials and fashion them into servants of mankind. To optimize their use of those materials, however, the designers and engineers need to know all about them. In other words they need property data for engineering materials.

Until the computer came of age, these experts either performed tests to obtain their own data or sought those data from a variety of printed or human sources. Given the expense, slowness and inconsistent quality of data obtained this way, it is not surprising that the computer was seen as a powerful component to be added to the information loop. Out of this grew the concept of "computerized" property data.

Significant efforts to deal with such data began about ten years ago and progress was amply illustrated at the recent Third International Symposium on the Computerization and Use of Materials Property Data[1]. Three days of presentations, coupled with demonstrations of working systems, showed that while much has been accomplished, much remains to be done.

The purpose of this paper is to sketch the current scene, primarily through a synopsis of the Symposium, but first it is necessary to outline some background material.

2. BACKGROUND

2.1. Definitions

An ordered collection of computerized property data for engineering materials is known as a "database". Databases may conveniently be divided into two broad categories: reference or source. "Reference" includes those databases that contain bibliographic citations or references to other information sources; "source" includes those databases that contain numeric, textual-numeric, full-text or image information[2]. In searching, therefore, a "hit" in a reference database points to a place where the desired information may be found whereas a "hit" in a source database contains the desired information item itself.

Thus, an engineering materials property database (MPD) is a source database. More specifically, it is an ordered collection of data items whose values (1) correspond to various large scale properties, parameters or attributes of practical materials and (2) are critically evaluated or validated by experts prior to their being included in the database.

Clearly, there are similarities between MPD and scientific numeric databases (SND)[3] but there are also significant differences. Whereas SND tend to deal with the fundamental, microscopic properties of elements and compounds, MPD tend to deal with the extensive, macroscopic properties of natural or fabricated substances which can depend upon the manufacturing process, geometry, use history and many other factors. This aspect of MPD will be amplified in Section 3.

2.2. History

It is generally felt that the history of computerized materials property data dates from an international workshop convened at Fairfield Glade, Tennessee in 1982[4]. This seminal meeting was followed by two others in Petten (1984[5], 1988[6]) and one in Schluchsee[7] (1985). As the field matured, international symposia were held in 1987[8] and 1989[9] with the most recent held September 1991 as mentioned above. In addition to these major gatherings, several discipline-oriented seminars were held in 1984-1986 to explore problems related to particular industries, the one pertaining to the aerospace industry being typical[10].

Progress was not limited to meetings, of course, and practical online prototype database systems, the Materials Property Data Network (USA) and the Demonstrator Programme (Commission of European Communities) began in 1984 and 1986 respectively.

It is fortunate that, from the outset of this activity, the need for international standards has been acknowledged. In this way cooperation has been fostered, duplication of effort avoided and the need to strike compromises, to satisfy strong factions with pre-existing vested interests, circumvented. Groups like the Technical Working Area 10 of the Versailles Project on Advanced Materials and Standards (VAMAS) and the Committee on Data for Science and Technology (CODATA) of the International Council of Scientific Unions foresaw the need for standards and pressed vigorously for them to be developed. Responding to this challenge, the American Society for Testing and Materials (ASTM) formed Committee E-49 in 1986. Formally designated the Committee on the Computerization of Materials Property Data, E-49 was given the task of developing standard classifications, guides, practices and terminology for building and accessing materials property databases. Only through such internationally accepted consensus standards can the quality and reliability of MPD be maintained and compatibility among different applications be assured in order to reduce costs and promote exchange.

3. PROBLEMS UNIQUE TO MPD

To understand why MPD have required such a vast effort and the number of practical databases is still relatively small, it is necessary to grasp why these data are more difficult to deal with than fundamental property data.

Fundamental properties, like for example the specific heat of pure aluminum or the crystal structure of salt, are an intrinsic property of that substance. Hence they tend to be constant in time, independent of the manufacturing process and independent of the direction or technique by which they are measured. Furthermore, everyone agrees the symbols "Al" and "NaCl" uniquely identify the substances in question and can reproduce those substances, and the measurement if need be, in their laboratory with no additional information. The corresponding entry in a database of fundamental properties in principle, therefore, need be little more than a numerical value assigned to field name.

Engineering properties, like for example the creep strength of steel or the fracture toughness of an aluminum alloy, are not intrinsic properties of those substances. They may change as the material is loaded and as it ages, they will certainly depend on the heat treatment used in manufacture, and the numerical values assigned to those parameters really have meaning only if cited in conjunction with the test method used to determine them. Furthermore an engineer in one country may know a particular alloy of steel as "Type ABC" whereas a colleague in another jurisdiction might know it as "Type XYZ". Thus, to be meaningful, entries for these properties in a database must be more than rudimentary numerical values; they must also include data about the data, otherwise known as "metadata", which may be defined as a set of data descriptors and other associated information that characterize the individual data values.

Clearly, these examples are simple in the extreme but should illustrate the inherent "fuzziness" of MPD. While the computer is essential for organizing and disseminating MPD, it is frustratingly "mindless" and rigid when it must be harnessed to deal with fuzzy entities. It is for this reason that so much effort has been expended in learning and agreeing how, in a computer environment, (1) to specify engineering materials unambiguously; (2) to determine what minimum set of data and metadata are required to define a given property; (3) to stipulate the quality of those data; (4) to convey that succinctly through uniform vocabulary to users with disparate backgrounds and interests.

An additional complication that must be faced by producers of MPD is that the quality of data needed varies with the objectives of the user. Thus, during conceptual design, the data need only be approximate but should cover a wide range of material classes so that all plausible candidates are considered. During preliminary design, range of coverage is no longer a factor but now higher precision and reliability are. Ultimately, in the final design stage, the best accuracy and precision are mandatory. Because scope of coverage and high accuracy are essentially independent goals, each attained only at considerable cost, compromises must be made.

4. PROGRESS TO DATE

As mentioned in the Introduction, the recent Third International Symposium on the Computerization and Use of Materials Property Data provides an excellent overview of the state of this activity. In reviewing the proceedings, it is convenient to use the categories used by the program organizers.

4.1. Standards and Data Representation

Because it is so costly to obtain and disseminate and so vital to a proficient manufacturing industry, materials information must be regarded as an international commodity. Harmonization of such basic entities as terminology, materials description, data representation and data quality is the key to sharing that commodity efficiently and economically.

Terminology, for instance, must attempt to cover comprehensively and accurately the vast range of materials (metals, polymers, refractories, composites, etc.), must cope with three different origins of nomenclature (the data source, the database producer and the user) and must admit hierarchical and synonymous thesaurus relationships enabling a "mindless" computer to expand, translate or link terms appropriately[11].

Lest anyone think that materials databases could be trivially produced by scanning printed works and feeding the results into a computer, it was shown that this process is anything but straightforward. A detailed account of the problems encountered in capturing the information in a major reference such as the Military Handbook 5E amply demonstrated the impracticality of attempting to convert such data collections to machine readable form solely by automatic means[12].

4.2. Standards and Database Development

In its relatively brief existence, Committee E-49 has already succeeded in producing five ASTM Standards[13]. Given that such Standards are only adopted after scrutiny and consensus of progressively wider segments of the ASTM membership, as a proposed standard proceeds from the initial drafting group to the Society level, this is indicative of the intense level of activity in this area.

E1407, the most recent Standard, relates to the quality and reliability of data, database system management, system capabilities and data security. The guidelines described are intended for both producers and distributors of MPD and are applicable to all delivery systems whether personal computer diskettes, CD-ROM disks, magnetic tapes, local area networks or public telecommunications networks[14].

To facilitate exchange of data between databases and between databases and applications, it is essential to develop a mutually agreed upon, machine independent format in which the data may be arranged. Without such neutral exchange formats, anyone wishing to acquire data from another source must first convert those data to a form recognized by his application. Repeating that exercise for every new data source is obviously an inefficient, error-prone procedure; hence the interest in neutral exchange formats. To date there exists no widely accepted format but some organizations have developed ones for their own use[15]. It seems likely, in fact, that a workable format may emerge as a subset of the formats being developed in the STEP program. STEP (Standard for the Exchange of Product Model Data) is an activity of the International Standards Organization aimed at developing a generic, computer-compatible means of describing manufactured goods[16].

4.3. Expert Systems and Materials Databases

As evidence that the field of MPD is maturing and interest is expanding to matters other than the data themselves, a complete session was devoted to the role of expert systems.

The user interface is one application receiving attention. It is proposed to use expert systems to assist a novice user in formulating a database query or to aid the user to appreciate the limitations of the data retrieved. Potentially, expert systems offer a means of handling "fuzzy" data, from both the query and display points of view, that cannot afford to be ignored. On the other hand, in the inevitable zeal to apply these tools, one must not neglect their inherent limitations[17].

Another application described was the prediction of behaviour, cracking of stainless steels in a hostile environment, in this instance, by interpreting field data in terms of an extensive knowledge base of rules and materials data[18].

In a dramatic live demonstration, one group demonstrated the use of expert systems as a tool for teaching students to exploit MPD in a creative, intuitive manner. Rather than bury the student in reams of tables and charts, the system encourages him/her to think in terms of so-called "Merit Indices", combinations of materials properties which, if maximized, maximize performance. Such a system almost automatically forces a designer to look at competitive materials thereby avoiding the tendency simply to use what has worked in the past and possibly leading to more innovative, optimized products[19].

4.4. Data Issues for Engineering Materials

While, as described above, work in developing standards for dealing with engineering materials in computer systems began only quite recently, other standards relating to materials have existed for years. These standards, having for example to do with testing, classification, designation and use, were produced without the computer in view and are not necessarily easily adapted to computerized systems. Furthermore, they have a certain inertia and status from years of acceptance; practitioners will therefore not readily abandon them in favour of a new scheme just because it is claimed to be more amenable to computerization. Designation schemes are a case in point. Existing schemes tend to be mnemonic, sacrificing comprehensiveness of detail for ease of use; designers of computerized schemes prefer designators to be comprehensive because an unintelligible series of characters is not a problem to a computer. Thus there is a certain tension between practical, working standards and standards ideally suited to computerization[20].

Advanced composite materials are being used to replace the more traditional aluminum alloys in aircraft, aerospace and naval applications. These materials are inherently difficult to characterize because their property data are especially sensitive to such factors as composition, configuration, processing and test methodology. Data for composites therefore require exceptionally carefully evaluation before being added to a database along with extensive, well-documented metadata[21].

Not all issues related to MPD are technical. If materials databases are to be successful, they must meet real needs and their perceived value must be such that a user is willing to pay a fair price to access them. Conclusions from the prototype online system (Materials Database Demonstrator Programme) of the Commission of European Communities (CEC) indicated, in fact, that trialists regarded data as having low value. It is evident therefore that potential clients must be educated to view technical information like MPD as a valuable resource, created through financial and intellectual investment. Directorate General XIII of the CEC is about to embark on a program to address these and related concerns[22].

4.5. Industrial Applications

The integration of materials selection into the engineering design process, in an in-house Materials Information (Database) System, was reported by one aerospace company as an important element in providing a competitive advantage. Until the next breakthrough in materials technology, manufacturers remain competitive chiefly by continually refining what they have done before. Improvements in the quality of MPD permit design margins to be narrowed and manufacturing cost savings to be identified. Design and development times and costs are reduced because fewer design iterations are required to attain a given set of objectives.

To enhance the utility of the System, provision is made to feedback experience gained in using a given material and to document the reasoning behind design choices. With this accumulation of design experience the updated database progressively becomes a more valuable company resource[23].

A role for MPD has also been found in structural integrity assessment programs. Numerous large-scale engineering complexes, such as steam generation or petro-chemical plants, operate under conditions in which the crucial properties of some materials diminish with service. Thus there is a need for databases containing information on alloys whose properties are degraded with long periods of stress and high temperature service. Incorporating such MPD with residual life models in an online, non-invasive monitoring system would give management real-time analysis of plant integrity along with an ability to forecast useful life time[24].

At the other end of the technological spectrum, NASA is developing a database of fracture mechanics properties of materials for use in fracture control analysis of space hardware. In light of the discussion in the first paragraph in sub-section 4.4 above, it is interesting to note that the database developers found it expedient to devise a specialized, "intelligent" (ie. non-mnemonic & essentially unintelligible to a human) identification code for their data[25].

4.6. Working and Prototype Systems

Tangible evidence of progress in MPD was provided by demonstrations of a number of commercial and prototype systems. (A list of participating organizations is given in the Annexe.)

That all except one of the demonstrations were microcomputer-based rather than online to a remote host, is somewhat indicative of the manner in which this field is evolving. (In this particular location, lack of easy access to telecommunications may have mitigated against more online demonstrations but the general trend is still evident.) Thus, quite reasonably and logically, MPD developers are currently concentrating on relatively homogeneous subsets of materials thereby serving a defined market and gaining experience without being engulfed by the complexities of dealing with a broad range of materials. The exception to this is the online Materials Property Data Network which already covers plastics, aluminum and steels with plans to broaden its scope.

5. CONCLUSIONS

It is evident that much has been accomplished since the Fairfield Glade meeting in 1982. What was then a proverbial gleam in the eyes of a few visionaries has grown into a maturing sub-field of the informatics industry. Databases now exist covering an impressive range of practical engineering materials; search systems vary from simple "look-up" to artificial intelligence-assisted queries; standards have been developed to describe engineering properties and data quality consistently.

Future work will certainly involve further efforts to develop standards for facilitating data exchange and to extend data recording formats to additional materials. Increasingly intelligent and helpful "front ends" will be perfected to assist users at all levels of expertise. It is highly probable that online and stand-alone systems will evolve in parallel, each addressing their respective markets.

Engineering students in the next century may well be as blasé about their materials data as the general public is now about their utilization of "ordinary" structures and devices. Thanks to their predecessors, these students should be free to major in design innovation rather than in data pursuit.

ANNEXE

Demonstrations of Computerized Materials Property Data Systems and Expert Systems

1. Achilles: Corrosion Expert System
   National Corrosion Advisory Service
   National Physical Laboratory, UK

2. BASF Plastics Materials Information Systems
   BASF plc, UK

3. Computerised Selection of Powder Materials for Structural Components
   MPR Publishing Services Ltd., UK

4. COMAR: International Databank for Certified Reference Materials
   Laboratoire National D'Essais, France & Laboratory of the Government Chemist, UK

5. Copper Select
   Copper Development Association Inc., USA

6. Engineering Materials Selector
   Department of Engineering
   University of Cambridge, UK

7. Header Database & FATDAC
   ERA Technology, UK & Failure Analysis Associates, USA

8. MatDB: Materials Information System
   ASM International, USA

9. MATEDS, Materials Technology Education System
   Royal Institute of Technology, Sweden

10. Materials Databases from NIST Standard Reference Data
    National Institute of Standards and Technology, USA

11. Materials Property Data Network
    STN International, USA

12. MATUS & Copper Data Disks
    Engineering Information Company Ltd., UK

13. MORPHS: Software for the Natural Rubber Formulary
    Rubber Consultants, UK

14. MTDATA: The NPL Databank for Metallurgical Thermochemistry
    National Physical Laboratory, UK

15. M-Vision - Composites
    PDA Engineering, USA

16. PAL: Expert System for Selecting Industrial Adhesives
    Permabond, UK

17. PERITUS Engineering Materials Database System
    MATSEL, UK

REFERENCES

1. Third International Symposium on the Computerization and Use of Materials Property Data, Downing College, Cambridge, UK; 9-11 September 1991; sponsored by ASTM Committee E-49 on Computerization of Materials Property Data and the UK National Physical Laboratory.

2. <u>Directory of Online Databases</u>, Cuadra/Elsevier, New York, 1991, Vol. 12, p. vii.

3. Wood, G.H. et al., "Canadian Scientific Numeric Database Service", <u>J. Chemical Inf. and Comp. Sci.</u>, 1989, 29, 118;

Wood, G.H., "Scientific Numeric Databases", Paper 5, AGARD Conference Proceedings No. 357, <u>The Application of New Technologies to Improve the Delivery of Aerospace and Defence Information</u>, (Technical Information Panel Specialists' Meeting, Ottawa, September 1983), AGARD-CP-357, 1983.

4. Westbrook, J.H. and Rumble, J.R., eds. "Computerized Materials Data Systems", Proceedings of the Materials Data Workshop, Fairfield Glade, TN, 1982, 133 pp.

5. Kröckel, H., Reynard, K. and Steven, G., eds., CEC "Workshop on Factual Materials Data Banks", Petten, The Netherlands, 14-16 November 1984,

Office for Official Publications of the European Communities, Luxembourg, 1985.

6. Büttner, P. and Kröckel, H. eds., "Report of a VAMAS Workshop on Standards for Materials Databanks", Petten, The Netherlands, 15-17 November, 1988, VAMAS Technical Report 4, 1989.

7. Westbrook, J.H., Behrens, H., Dathe, G., and Iwata, S., eds., "Materials Data Systems for Engineering", Proceedings of a CODATA Workshop held at Schluchsee, Germany, 22-27 September, 1985, Fachinformationszentrum Karlsruhe Gmbh, Germany.

8. Glazman, J.S. and Rumble, J.R., eds., "Computerization and Networking of Materials Data Bases", Proceedings of the First International Symposium on Computerization and Networking of Materials Property Data Bases, 2-4 November 1987, ASTM STP 1017, ASTM, Philadelphia, 1989.

9. Kaufman, J.G. and Glazman, J.S., eds., "Computerization and Networking of Materials Databases: Second Volume", Proceedings of the Second International Symposium on the Computerization of Materials Property Data, 29 Nov.- 1 Dec. 1989, ASTM STP 1106, ASTM, Philadelphia, 1991.

10. Westbrook, J.H. and McCreight, L.R., eds., "Computerized Aerospace Materials Data", Proceedings of a Workshop on Computerized Property Materials and Design Data for the Aerospace Industry, El Segundo, CA, 23-25 June, 1986., American Institute of Aeronautics and Astronautics Inc., 1987.

11. Westbrook, J.H., "Terminology Standards for Materials Databases", Third International Symposium on the Computerization and Use of Materials Property Data, Cambridge, UK, 9-11 September, 1991, Proceedings to be published by ASTM.

12. Grattidge, W. et al., "Problems in the Computerization of a Printed Reference Work on Materials Data", Third International Symposium.

13. E1313-90, E1314-89, E1338-90, E1339-90, E1407-91, Annual Book of ASTM Standards, ASTM, Philadelphia, 1991 (Published annually).

14. Kaufman, J.G., "Quality and Reliability Issues in Materials Databases: ASTM Committee E49.05", Third International Symposium on the Computerization and Use of Materials Property Data, Proceedings to be published by ASTM.

15. Cverna, F.A. et al., "An ASCII File Format for Materials Properties Database Import and Export"; Vinard, D.R. et al., "Use of Z 99-001 as a Neutral Exchange Format for Saint-Gobain's Materials Databank", Third International Symposium.

16. Rumble, J.R. Jr., "The STEP Model of Materials Information", Third International Symposium.

17. Vancoille, M.J.S. et al., "Probing the Boundaries of Knowledge Based Systems", Third International Symposium.

18. Komai, K. et al., "Development of Diagnostic Expert System for Environmentally Assisted Cracking (EXENAC) and Importance Evaluation of Knowledge in Inference", Third International Symposium.

19. Cebon, D. and Ashby, M.F., "Materials Selection for Mechanical Design", Third International Symposium.

20. Reynard, K.W., "Standards for Materials Databases - National and International Programmes - Do They Provide for Users Needs?", Third International Symposium.

21. Newton, C.H., "Data Analysis and Evaluation of Advanced Composite Materials", Third International Symposium.

22. Swindells, N., "Information Engineering, Adding Value to Data", Third International Symposium.

23. Newley, R.A., "The Integration of Materials Information into Engineering Design", Third International Symposium.

24. Jeffrey, C.M. and Bullough, C.K., "In-Service Structural Integrity Assessment Using Computerized Materials Models and Data", Third International Symposium.

25. Lawrence, V. and Forman, R.G., "Structure and Applications of the NASA Fracture Mechanics Data Base", Third International Symposium.

FACILITATING THE TRANSFER OF SCIENTIFIC AND TECHNICAL INFORMATION WITH SCIENTIFIC AND TECHNICAL NUMERIC DATABASES

H. Haller

Defense Technical Information System
Office of Information Systems and Technology
Cameron Station
Alexandria, VA 22304-6145
United States

SUMMARY

The Defense Technical Information Center (DTIC) provides services primarily to librarians and technical information specialists. In an effort to better serve engineers and scientists, the end users, DTIC conducted a technology assessment of users and developers of scientific and technical numeric databases. DTIC's Department of Defense Gateway Information System (DGIS) provides the access mechanism to databases and the Multi-Type Information and Data Analysis System (MIDAS) will provide the capabilities to process bibliographic information and numeric data.

I. INTRODUCTION

The Defense Technical Information Center's (DTIC) mission includes facilitating the transfer of scientific and technical information within the Department of Defense (DoD). Librarians and technical information specialists represent a major component of DTIC's customer base who often serve as intermediaries between DTIC and scientists and engineers, the end users of the information. DTIC provides several online systems for their use in providing technical information to researchers within their organizations. The first of these, the Defense RDT&E Online System (DROLS) has provided classified and unclassified access to databases generated by DTIC since 1974. The most recent system, the Department of Defense Gateway Information System (DGIS), provides access to over 700 databases generated by government and commercial entities.

DGIS

Dialing into DGIS, users can automatically connect to and search these remote services, download the search results onto the DGIS host computer at DTIC and then process the results using DGIS analysis and report generating tools and utilities.

DGIS has provided many searching and processing functions needed by users of bibliographic information. These include performing multi-tasking (i.e. running multiple programs concurrently) eliminating duplicate citations, and providing both menu and command driven systems as chosen by the user. However, there are some functions that at this time might be better performed remotely, at the customer's site. These include retrieving and processing the full text of a record and marking citations for retrieval. Desktop workstations provide the capability for improved user interfaces. Improved user interfaces -- what the user sees -- can include multi-tasking in a windowing environment, pop-up windows, and hypermedia.[1]

Presently, the DGIS interface is character-based. Personal Computer (PC) users with a graphical user interface rather than a character user interface are more productive, more accurate, and less frustrated in their work, regardless of the individual's microcomputer experience.[2] DGIS' connectivity to a variety of information services containing a wide range of subject matter has implications which go beyond accessing varied information. With improved interfaces to facilitate ease of use, the DGIS user base can expand from the traditional information intermediary to include the end user; the engineer and scientist can be served directly.

A PC user dialing in to DGIS must run terminal emulation software compatible with a UNIX host. They run a terminal emulation program to access DTIC's UNIX system. This allows PC users to continue to run their favorite applications. However, the customer cannot easily make full use of the UNIX multi-tasking capabilities nor does the user interact in a windowing environment with a graphical user interface because of DGIS' character-based interface.

The gateway concept has created a dichotomy. On the one hand, DGIS customers have dial-in capability over a wide area network to connect to varied resources and use the disk space and processing tools at the host computer. On the other hand, they need a graphics based interface. For most of our customers with their current telecommunications, connecting over a wide area network to a system with a windowing environment and a graphical user interface is presently impractical. Our customers lack high speed lines for intensive processing required to send graphic instructions. Until the availability of more rapid telecommunications, a possible solution is to offload some of the centralized processing from the host to the remote system and use DGIS to gateway to other information systems. The Multi-type Information and Data Analysis System (MIDAS) is an attempt to do just that.

MIDAS

DTIC customers are a disparate group with a variety of computer environments, information resource needs, and information processing needs, not unlike other computer users. That is, they use a variety of hardware platforms -- IBM-compatible and Apple Macintosh PCs, UNIX-based workstations, minicomputers, and mainframes with varying operating systems. Therefore, DTIC must develop applications that are interoperable, that is, not machine nor operating system dependent. Regardless of the hardware or operating system used, the application should have the look and feel with which the user is most familiar. However, DTIC cannot incur the expense of copying and recompiling programs onto different machines. Programs developed under the X Window standard, a graphics and network protocol, eliminate many of these portability issues. This prototype will show X Windows-based UNIX programs running on PCs and sharing information between DOS, Mac and UNIX programs.

Customers' data and information needs vary as widely as do their computer systems, networking, and accessing capabilities. Their data resources vary in media and distribution formats. Besides bibliographic, there are other database types like those containing full text and numeric information. With the influx of multi-media platforms, audio and image databases will soon become more prevelant. Also, data and information usage, and therefore the processing and analysis tools and utilities required, vary among intermediaries and end users. So do the kinds of data and information needed whether it is tabular data, full text of articles, audio or graphics.

The project, Multi-type Information Data Analysis System (MIDAS) will respond to these issues. Developing a MIDAS prototype will demonstrate DGIS with a graphical user interface, a remote connection to a gateway of information from a local workstation environment, incorporating interoperability, accessing different database types and processing other types of data. In developing a MIDAS prototype, we make the assumptions that the workstation and PC market (UNIX-DOS-Mac) are the typical computing environments within our user community.

The prototype will be developed on a Sun workstation and will demonstrate connectivity using the Sun as a server to run applications and a PC as a terminal. Of the largest number of new workstation buyers, 24% of the people are former PC users. The biggest interest in workstations is from the engineering and technical community.[3]

Since DTIC is in the business of transferring scientific and technical information to the Department of Defense (DoD) community, the prototype will include scientific and technical numeric databases, primarily materials properties.

Scientific and Technical Numeric Databases

DGIS provides some solutions to the present day processing of bibliographic information by connecting to different database types all over the world. Through DGIS, DTIC can actively assist end users in accessing and processing not only bibliographic information, but other types of information, such as numeric data.

Numeric databases are collections of information and data. They contain both data and metadata or textual information relating to the data. There are many different kinds of numeric databases. Scientific, technical and engineering databases comprise the second highest subject category of all numeric databases next to business databases.[4]

As DTIC's mission is to facilitate the transfer of scientific and technical information, this is our primary numeric database subject area of interest. Within that category, the many types of online sources include chemical, physical, and materials properties databases. Materials properties is an area in which many DoD engineers and scientists have an interest, as products from aircraft and missiles to clothing are produced from materials. Their interests lie in the effect that conditions such as stress, strain, and temperature have on a variety of materials. To determine how numeric data can be processed in MIDAS, DTIC contacted engineers and scientists.

II. TECHNOLOGY ASSESSMENT--USER NEEDS RESEARCH

DTIC sponsored a Scientific and Technical Numeric Database Technology Assessment to identify individuals who have a need for new services that DTIC should provide; to identify additional needs (resources and tools) these individuals have; and to assist in responding to those needs. The assessment concentrated on work practices, media, and time spent in information seeking activities. The study measured bibliographic information and numeric data searching frequency, primary sources, user satisfaction with sources, preferred media, materials and materials properties interest, accessibility and computer capabilities and requirements.

The assessment directly involves end users--engineers and scientists. With 86% of respondents consisting of engineers, the remainder was evenly split between scientists, technical information specialists and/or librarians and others. DTIC mailed about 20,000 assessment forms to individuals in the DoD and industry between October 1989 and January 1990 and received 1,601 responses. Roughly, 73% responding were DoD employees, 22% were from industry and the remaining 5% were other government agencies and universities. Results are presented below, grouped by major categories.

Work Practices

Project management, research, and testing were the most frequently specified job functions. Most respondents indicated they rely on paper as the medium for acquiring and using scientific and technical information. But, a significant minority use floppy disk. Nearly one-third spend over 25% of their time in information seeking activities. However, nearly two-thirds spend the same amount of time actually using scientific and technical information. See Appendix for Tables 1 and 2.

Information Needs--Bibliographic Information

1. Frequency of Searching for Bibliographic Information

Approximately 25% of the respondents frequently search for bibliographic information (source citations).

Yes, search: 419	No, do not search: 1,123

2. Primary Sources

Responses to the request to list sources revealed that most of the respondents rely on five sources for bibliographic information. These are listed in the Appendix in Table 3 in order of most used to least used.

3. Level of Satisfaction with Current Source

The question was designed to measure the level of satisfaction participants experience with their current bibliographic sources regarding accessibility, ease of use, presentation, quality, user support and price. The majority of responses indicate that participants rank their data sources as "good" to "needs improvement". See Table 4 in the Appendix.

Information Needs--Numeric Data

1. Frequency of Searching for Materials Properties Numeric Data

Regarding the frequency of searching for numeric data, just under 50% of those responding search for data.

Yes, search for data: 751 No, do not: 811

2. Primary Sources

Participants listed handbooks most frequently as their primary source of materials properties numeric data. Other sources are listed in Table 5 of the Appendix.

3. Level of Satisfaction with Current Source

The assessment was designed to measure the level of satisfaction participants experience with their current data sources regarding accessibility, ease of use, presentation of data, quality of data, user support and price. The majority of responses indicate that participants rank their data sources as "good" to "needs improvement". Table 6 in the Appendix, summarizes the results.

Preferred Data Sources and Media

1. Preferred Data Sources

This question tested six data sources for their importance to the participants and the preferred medium for working with each of the sources. The overall observation is that engineering handbooks, military handbooks, internal laboratory reports and journals are most important to these participants and that the vast majority of the respondents currently work in a paper medium. See Table 7 in the Appendix.

2. Preferred Media for Materials Properties

When asked to indicate media preference for working with materials properties, the respondents provided the choices shown in Table 8 in the Appendix. Participants' preference for the future is to work with information sources in a computerized environment reducing their dependency on paper.

Defining Materials and Materials Properties Data Interest and Accessibility

Participants indicated their level of interest and degree of accessibility in procuring information. Materials and materials properties which received the most responses as "critical" to respondents' work are listed below.

Materials	Materials Properties
Alloys	Chemical
Metals	Mechanical
	Physical
	Thermal

The materials and materials properties receiving a significantly high level of interest response but were considered to be either"accessible with difficulty" or " inaccessible with current resources" were: Carbon Matrix and Polymer Matrix Composites, and Optical and Thermoradiative. See the Appendix, Table 9 for the summary results.

Uses of Materials Properties Data--Current and Future Uses

Responses to this question show that the respondents consider computer simulation the most significant increase in their use of materials properties data in the future. The need for this information in engineering calculation, materials selection, materials engineering, structural design, product testing and quality assurance decrease in the future.

Computer Usage

Most of the participants responding to the question "Do you use a computer in your work?" gave a positive response; 1,460 answered yes and only 114 answered no.

1. Hardware

For this group of respondents, the IBM compatible PC environment received the most responses. Table 10 in the Appendix lists hardware in decreasing order of use.

2. Graphics

Most (80.7%) of the participants responding indicated that they have graphics capabilities, i.e. graphics cards and monitors.

3. Modem

With regard to data communications capability via modem, approximately 45% of the respondents have a modem.

4. Software

The types of software functions were surveyed and results are listed in Table 11 of the Appendix.

In a separate question with a free response format, participants were asked to list other software functions that are important to them. These include: analysis programs, database, graphics, spreadsheet, and word processing.

III. INTERVIEWS--FUNCTIONAL DESCRIPTION

We developed a profile of end users based on the tabulation of responses for two purposes: write a functional description of the MIDAS prototype and identify important sources of information and data required. The profile includes a representative sample of individuals with the following attributes:

o job functions include: project management, research, and testing

o acquisition and use of scientific and technical information consumes 25-50% of the workday

o a critical interest in materials including metals, alloys, and composites

o an interest in materials properties that are accessible with difficulty including mechanical, physical, thermal, chemical or inaccessible including thermoradiative and optical properties

o search numeric data and bibliographic literature regularly

o diskette is the preferred media for numeric data sources

o software function needs include but are not limited to: modeling, in-house database, graphic comparison, statistical analysis, unit conversion, CAD/CAM and downloading capabilities.

Based on the above-mentioned profile, we selected individuals in the Washington, D.C. area to interview to obtain greater detail regarding work products, data sources and needs. Areas of interest include: current and future uses of materials data, current sources of materials data, sources used in the past and abandoned, means to acquire data, complaints about the data sources, accuracy requirements, and detailed work environment descriptions by individual and organization (technology, computer environment, processes of data input and output, existing hardware/software tools and formats, existing programming facilities available).

Interview Responses

The engineers and scientists interviewed have unique data applications and often use different resources. This aspect of the project is continuing. However, we have some preliminary findings regarding resources, computer capabilities, data usage, and software functions.

1. Resources

Many users must perform literature searches primarily for the purpose of determining whether similar work has been performed in the past or how certain problems have been solved by other organizations. They do not conduct their own searches but rather request their libraries to assist. Many complain that it takes too long to get results and that often they must justify why they need the information because of budgetary cutbacks.

Most offices are comprised of engineers and scientists who either are long-term employees or have very recently come to work with the government. Nearly all individuals interviewed have indicated that they have a need to share information with others either inside their own organization or outside with other agencies and vendors. Within an organization, there is a wealth of knowledge to tap. Often these are the same individuals who have points of contact outside the organization. Most of the interviewees rely heavily on communicating with other experienced, well-established engineers and scientists for information.

Some individuals feel they can not afford access to several desirable online services. Most engineers interviewed have at least periodic need to review military specifications, engineering handbooks, and test methods endorsed by American Society for Testing and Materials (ASTM) and other standards organizations.

By and large, most work products are not shared beyond the immediate purpose for which the information and data were gathered and generated. Final reports are often stored within the organization in hardcopy with the individual author retaining the report on disk. A few individuals do maintain a database of information so reports are easily accessible by others for later use.

2. Computer Capabilities

Most people interviewed are using IBM-compatible PCs that have been upgraded from an Intel 286 to the 386 processing chip. Within an organization, most of the individuals are either on a local area network or are in the process of being networked. Many have network access to mini or mainframe computers on-site. Usage for such systems may or may not incur charges. Often there is dial-out capability from these systems which is rarely used.

Many engineers use testing equipment with embedded computer software programs to generate tables of data, graphs and electronic photographs. When compiling information and data from these systems, there is little data sharing which results in re-entry into the PC for report generation.

The individuals who most frequently use computers prefer to have their information and data in some computerized format. The individuals who prefer paper tend not to be computer users or are more comfortable with more traditional scientific and technical practices. Most of the interviewees do their own programming. They feel they best know their own needs and that they are most capable of developing a useful system.

Many use off-the-shelf software packages for putting together their final products. Their products consist of analysis, evaluation and recommendation reports. Most individuals interviewed felt they lacked the time to learn new packages and preferred the simple, easy to learn software. They cannot afford a long learning curve. Most frequently they cut and paste graphs, tables and text accessible from different computer systems and even photocopy pages from books.

3. Data Usage

Many of these individuals generate a lot of their own data, primarily in testing and research. Some interviewees are involved in the actual design, materials selection and definition of system specifications. Others identify problems with components and verify that the components met original specifications. Typical data usage includes stress-strain analysis, failure analysis, vibrations analysis, environmental testing, prototype testing, and testing of developed, operational systems. Some interviewees look for alternative materials, some have to identify the materials, and for some, the composition is irrelevant. Most are limited to specific materials of interest but many feel they must keep abreast of other new materials and technologies.

4. Software Functions

Most interviewees use many different software packages for gathering information and data for presentation purposes. However, Harvard Graphics was the most frequently mentioned package. Many individuals expressed a desire for unit of measurement conversion. Most of those interviewed still report results in English and not in metric units, and conversion is either done manually or through some relatively simple software programs. The most frequent desire for modeling data was finite analysis. There are certain favorite models that these individuals use and the most significant problem is in tracking data generated from the models. Some individuals only perform modeling with no testing on site. Most expressed a desire to have graphic comparison capabilities in either two or three dimensions. Often they have to re-enter data into these packages. Some interviewees have in-house databases but others who do not, want simple systems for tracking final reports and querying specific materials data contained in those reports.

IV. TECHNOLOGY ASSESSMENT--DATABASE DEVELOPERS RESEARCH

The purpose of the other phase of the technology assessment, Database Developers Research, was designed to identify designers, developers and distributors of scientific and technical numeric databases. Using a data call, we collected information regarding participants, the database and its content, the producer and the distributor points of contact, database operations, system capabilities, and requirements for user access. The data call encompassed any type of scientific and technical numeric database and was not limited to materials properties.

Distribution

The data call was mailed to 650 potential database designers, providers and distributors (after initial phone contact) between March and June 1990. These organizations were identified from a variety of information resources. The type of organization represented by these 650 individuals were classified as follows:

DoD	215
Foreign	107
Industry and Other Govt	328
Total	650

Responses

Of the 650 data calls sent, 231 were received for an overall response rate of 35.5%. The organization affiliation of the respondents is shown in the following table.

Organiz Affil	All Resp	Mat. Prop. Num. DB	Other Sci Tech DB
Civ govt	15	5	4
Commercial	69	32	12
Foreign	26	17	1
DoD	102	17	55
University	19	5	3
Total	231	76	75

The databases described in the responses fall into the following categories:

Materials Properties Databases	76
Other Sci/Tech Numeric Databases	75
Negative Responses	80
Total	231

For materials properties numeric databases, the commercial sector had the largest number of responses, 32 (42.1%). Foreign and DoD responses each accounted for 17 (22.4%) of the materials properties numeric databases reported. Civilian government and universities each accounted for 5 responses (6.7%).

For other types of scientific or technical numeric databases, the DoD had the largest number of responses, 55 (73.3%). The commercial sector was second, with 12 databases (22.4%). Civilian government accounted for 4 responses (5.3%), universities for 3 responses (4.0%) and foreign organizations for 1 (1.3%).

Data Call Updating

The data call demonstrated that numeric database design and distribution is a dynamic processs. Developers frequently change addresses, make significant modifications to database capabilities or contents, and even declare a database obsolete. A number of the databases represented in DTIC's "Directory of Resources", a directory accessible on DGIS, have been declared obsolete or the developers have moved on to other activities, leaving no one responsibile for the database. There is value in identifying even databases that are obsolete. Knowing that data were collected and perhaps accessible could provide the end user with a potential resource.

Of the 151 scientific or technical numeric databases represented in the responses, 33 (21.9%) are also represented in DTIC's existing Directory of Resources. The remainder, 118 (78.1%) are databases new to DTIC's directory. To maintain a current and accurate directory, the frequency of change demands an update cycle of at least once a year.

Range of Databases
The group of scientific and technical databases identified cover a broad range of subject matter from medical personnel records to oceanic temperature gradients. Individually, they cover a specific area of interest designed for a particular project or select group of users. Often, the database designers have not identified potential users beyond those for whom the database was specifically designed, lacking funds or the mission.

Database Media and DGIS
Though online access or distribution for databases in general is the most prevelant medium,[5] for this particular collection of scientific and technical numeric databases, the most frequently used medium is magnetic disk.

DGIS, with its gateway capabilities to directly access a wide range of databases would be of great value, offering many more users access to specialized data. However, this could require a significant investment, depending on the number of resources to make available and on the medium especially if it is not accessible to users online. The extent of the investment cannot be estimated without information regarding which databases to make available.

Directory on Diskette
An effective approach in selecting databases to distribute or make accessible would identify databases that address the largest and most critical segments of the user community. Based on critical sources and materials and materials properties needs identified in the needs assessment, a more detailed evaluation of data call resources would identify those that are invaluable to the engineer and scientist and any gaps between data needs and the resources available.

As a result of this phase of the assessment, a prototype directory on diskette was developed. The purpose was two-fold. One, to determine how useful the databases developed are to the end user. And, two, to distribute to the user the data call information collected in an easy-to-use product. The data call responses were input to a dBASE IV database. Using the program, Clipper, we created an executable program, that is a stand-alone executable file that can be invoked directly from DOS on a PC. Therefore, the directory does not require any additional software to operate. The resulting product is a menu-driven system with pop-up windows. A user can query, display and print results from the database directory.

Individual engineers and scientists who were interviewed for the numeric processing tools are also beta testing the diskette directory. The information we obtain from them will include: usefulness of directories of resources and the media; usefulness of the resources identified and others that should be included; the completeness of the information; the directory's ease of use; and any other comments. This aspect of the project is continuing.

Numeric Processing
As discussed earlier, one of the goals of this project is to investigate how scientists and engineers use scientific and technical data to help us design tools for manipulating data. The responses to the data call show that approximately 37 percent of the databases (56 of 151) provide data manipulation or analysis capabilities. The most common are plotting or other type of graphical output and unit conversion.

V. CONCLUSION

DTIC's gateway system, DGIS, will become the vehicle by which a wide range of information and data services will be accessed by intermediaries and end users. MIDAS will provide a complete processing environment for both bibliographic information and numeric data. Based on the Scientific and Technical Numeric Database Technology Assessment, DTIC has a much more thorough understanding of engineers and scientists information and data resource needs and their computing environment. Also, we have identified scientific and technical numeric databases throughout government and industry. Based on detailed interviews with a select, representative sample of those needs assessment respondents, users of numeric data want and need many capabilities that MIDAS could provide. We are writing a functional description to provide the basis of the design of software functions. Also, we are obtaining comments from the community regarding the usefulness of databases contained in the prototype directory.

The MIDAS prototype is a first step towards handling a variety of information and data in varying formats. In the future, information seeking activities will be distributed to an individual's workstation. DTIC has already provided the mechanism to access *information* sources, DGIS, but the next step, a complete information and data analysis and processing environment, has just begun.

APPENDIX

Table 1: Primary Media

Medium	Number of Responses	Percent of Respondents
Paper	964	61.6
Floppy Disk	430	27.5
Online	159	10.2
Other *1	11	0.7

Total 1,564

*1 Others included CD-ROM, magnetic tape, Microfiche/microfilm, autocad, CPU, mylar and verbal.

Table 2: Amount of Time Spent Acquiring and Using Scientific and Technical Information

% of Time Acquiring	% Resp	% of Time Using	% Resp
1-10%	38.9	1-10%	21.4
11-25%	20.3	11-25%	19.1
26-50%	22.2	26-50%	32.0
51-100%	9.6	51-100%	27.5

Table 3: Primary Sources for Bibliographic Information

Source	Number of Respondents
Libraries	75
DTIC	70
Dialog	58
Journals/open literature	49
Chemical Abstracts	31

Table 4: Level of Satisfaction with Aspects of Bibliographic Information Sources

	Excellent		Good		Needs Improvement		Unsatis-factory		Total Response
Accessibility	132	16.2%	347	42.7%	263	32.3%	71	8.7%	813
Ease of Use	105	12.9%	390	47.9%	265	32.6%	54	6.6%	814
Data Presentation	66	8.3%	396	49.9%	279	35.2%	52	6.6%	793
Quality of Data	75	9.3%	424	52.6%	254	31.5%	53	6.6%	806
User Support	67	9.2%	298	40.9%	266	36.5%	98	13.4%	729
Price	93	13.3%	306	43.8%	220	31.5%	79	11.3%	698

Table 5: Sources of Materials Properties Numeric Data

Sources	Number of Respondents
Handbooks (unspecified)	263
ASM	109
Journals/Literature references	100
Military data sheets/specifications	96
Manufacturing catalogues, handbooks	93
Military handbooks	55
Technical papers/reports or DTIC	48

Table 6: Level of Satisfaction with Aspects of Numeric Data Sources

	Excellent		Good		Needs Improvement		Unsatis-factory		Total Response
Accessibility	156	12.9%	524	43.2%	451	37.1%	83	6.8%	1214
Ease of Use	108	9.0%	494	41.1%	492	41.0%	107	8.9%	1201
Data Presentation	75	6.5%	525	45.4%	478	41.3%	79	6.8%	1157
Quality of Data	99	8.5%	578	49.4%	413	35.3%	80	6.8%	1170
User Support	45	4.1%	357	32.8%	485	44.5%	203	18.6%	1090
Price	126	13.1%	403	42.0%	301	31.4%	129	13.5%	959

Table 7: Importance of Key Materials Properties Data Sources

Source	Always		Sometimes		Never		Total
Engineering Handbooks	953	65.8%	405	28.0%	90	6.2%	1448
Internal Laboratory Reports	432	32.1%	636	47.2%	279	20.7%	1347
Journals	417	30.6%	675	49.6%	269	19.8%	1361
Military Handbooks	590	43.9%	509	37.9%	244	18.2%	1343
Product Specifications	347	54.7%	218	34.4%	69	10.9%	634
Technical Reports	275	47.7%	244	42.4%	57	9.9%	576

Table 8: Preferred Media

Medium	Number of Responses	Percent of Responses
Floppy/Hard Disk	698	30.8
Paper	638	28.2
On-line	490	21.7
CD-ROM	247	10.9
Microfiche/film	151	6.7
Magnetic tape	38	1.7

Table 9: Materials and Materials Properties Interest and Accessibility

	Interest				Accessibility				
	Critical	Useful	Not Important	Total Response	Easy Access	Access Difficult	Currently Inaccessible	Total Response	Difference
Material Groups									
Alloys	656 43.0%	543 35.6%	326 21.4%	1525	543 40.6%	647 48.4%	146 10.9%	1336	189
Ceramics	239 16.6%	533 36.9%	672 46.5%	1444	174 13.9%	750 59.9%	329 26.3%	1253	191
Composites									
Carbon Matrix	267 17.9%	560 37.6%	664 44.5%	1491	154 12.2%	724 57.3%	385 30.5%	1263	228
Ceramic Matrix	199 13.8%	493 34.3%	747 51.9%	1439	135 11.1%	678 55.9%	400 33.0%	1213	226
Metal Matrix	289 19.7%	558 38.0%	622 42.3%	1469	166 13.5%	698 56.7%	368 29.9%	1232	237
Polymer Matrix	296 20.5%	531 36.8%	614 42.6%	1441	183 14.7%	703 56.6%	357 28.7%	1243	198
Compounds	182 13.1%	583 41.9%	628 45.1%	1393	183 14.8%	767 62.1%	285 23.1%	1235	158
Metals	727 48.4%	501 33.4%	274 18.2%	1502	569 42.0%	661 48.8%	124 9.2%	1354	148
Polymers	285 20.9%	616 45.3%	460 33.8%	1361	162 13.4%	759 62.8%	288 23.8%	1209	152
Materials Properties									
Chemical	460 32.7%	570 40.5%	377 26.8%	1407	299 23.3%	769 59.9%	215 16.8%	1283	124
Electrical	281 20.5%	593 43.3%	497 36.3%	1371	224 18.9%	756 63.9%	203 17.2%	1183	188
Electronic	202 14.8%	523 38.4%	638 46.8%	1363	174 15.4%	700 62.0%	255 22.6%	1129	234
Magnetic	190 14.1%	538 40.1%	615 45.8%	1343	137 12.0%	711 62.3%	294 25.7%	1142	201
Mechanical	1 038 68.8%	305 20.2%	165 10.9%	1508	507 37.6%	737 54.6%	106 7.9%	1350	158
Optical	209 15.6%	430 32.2%	698 52.2%	1337	132 11.3%	684 58.8%	348 29.9%	1164	173
Physical	822 56.7%	438 30.2%	189 13.0%	1449	368 28.3%	809 62.2%	123 9.5%	1300	149
Thermal	685 46.8%	569 38.9%	210 14.3%	1464	266 20.2%	869 65.9%	183 13.9%	1318	146
Thermoradiative	154 11.6%	455 34.2%	721 54.2%	1330	98 8.6%	631 55.7%	404 35.7%	1133	197

Table 10: Hardware

Type of Computer	Number of Responses
PC - IBM compatible	1,231
Mainframe with CRT	456
Mini with CRT	261
PC - Macintosh	242
Graphics Workstation	198

Table 11: Software - Current Status and Needs

Software Function	Currently Have	Need
CAD/CAM	513	230
Downloading	483	207
Graphic compare	352	346
Inhouse database	557	355
Modeling	383	357
Statistical Analysis	463	343
Unit Conversion	195	299

[1] Reva Basch, "Database Software for the 1990s and Beyond--Part I: The User's Wish List," Online, (March 1990): 17-24.

[2] Lisa Day-Copeland, "Clinical Research Finds PC Users Are More Productive with GUI Interface," PC Week, 23 July 1990: 142.

[3] Robert D. Hof, "Where Sun Means to be a Bigger Fireball," Business Week, 15 April 1991: 73-74.

[4] Martha E. Williams, "The State of Databases Today: 1991; Forward," Computer-readable Databases: A Directory and Data Sourcebook, 7th edition.

[5] Ibid.

BRINGING DOWN THE BARRIERS TO INFORMATION TRANSFER

Technical Information Panel Specialists' Meeting
Madrid, Spain 8th—9th October, 1991

LIST OF PARTICIPANTS

Dr P.M. AGUADO BENEDI	Oficina Transferencia Investigacion, Baltasar Gracian 1, 50005 Zaragoza, Spain
Ms M.D. ALCAIN	Documentalista, ISOC, C/Pinar 25, 28006 Madrid, Spain
Mrs P. ALCALA	Director, University Library, Universidad de Alcala de Henares, c/Colegios 2, 28801 Alcala de Henares, Spain
Mrs M. ALVAREZ	Head, Technical Information, AMPER, S.A., C/Torrelaguna 75, 28027 Madrid, Spain
Mrs R. ARAMBURU	GMV, S.A., Issac Newton s/n, P.T.M., Tres Cantos, 28760 Madrid, Spain
Mr D. ARCHAMBAULT†	LIFO Université d'Orléans, B.P. 6759, 45067, Orléans Cedex 2, France
Dr J.H. ASHFORD	Ashford Associates Ltd, 72 Harrow Lane, Maidenhead SL6 7PA, United Kingdom
Mr J. BAJO ESTEBAN	Manager, CEMI, Conde Duque 9, 28015 Madrid, Spain
Dr J.-C. BASSANO†	Directeur de l'I.U.T. Informatique, LIFO, B.P. 6759, 45067 Orléans Cedex 2, France
Ms A. BENITO DE BENITO	Hortaleza, 56, 4, B, 28004 Madrid, Spain
Mr R. BERNHARDT*	GMD F 4, Institut für Integrierte Publikations-und Informationssysteme, Postfach 10 43 26, 6100 Darmstadt, Germany
Mr C.J. BIGGER*	Chief Librarian, GEC Marconi Research Centre, West Hanningfield Road, Great Baddow, Chelmsford, Essex CM2 8HN, United Kingdom
Mr W. BLADOS	c/o NASA Scientific and Technical/Information Program, NASA Headquarters, (Code NTT), Washington DC 20546, United States
Ms P BLANCO MUNOZ	Ministerio de Cultura, Servicio de Documentacion, Plaza del Rey 1, 28071 Madrid, Spain
Mr M. BRANDRETH*	Chief, Policy, Planning & Systems, Canada Institute for Scientific and Technical Information, National Research Council of Canada, Ottawa, Ontario K1A 0S2, Canada
Lt Col. H. BRAUN*	Dokumentations-und Fachinformations zentrum der Bundeswehr, Friedrich-Ebert Allee 34, D-5300 Bonn 1, Germany
Ms R. CABEZUELO	Documentalista, CEDISMAR, C/Juan Bravo 44, 28006 Madrid, Spain
Prof. J.A. CARVALHO	Universidade da Beira Interior, Rua Marques D'Avila e Bolama, 6200 Covilha, Portugal
Ing. Gen. F. CHEVALIER*	Directeur, C.E.D.O.C.A.R., 00460 Armées, France
Dr rer. nat C. VON CONSBRUCH*	Fachinformationszentrum, Karlsruhe GmbH, D-7514 Eggenstein-Leopoldshafen 2, Germany
Dr Ing. A.M.R. CORREIA*	L.N.E.T.I., Director, C.I.T.I., Azinhaga dos Lameiros, 1699 Lisboa, Codex, Portugal
Ms G. COTTER*	Director, Scientific & Technical Information Division, NASA Headquarters, (Code NTT), Washington DC 20546, United States
Mrs E. CURRAS	Calle O'Donnell 6, 28009 Madrid, Spain
Mr C. CZINCZENHEIM	Chef de Service, Dassault-Aviation, DGT/Bibliotheque Technique, 78 quai Marcel Dassault, 92520 Saint Cloud, France
Mr E. DE LA FUENTE GONZALEZ	Crisa P.T.M., c/Torres Quevedo s/n, Tres Cantos, Colmenar Viejo, 28760 Madrid, Spain

† Author
* Panel Member

Dr A. DEL REY*†	I.C.Y.T. (CSIC), Head U.E.I. Information Retrieval, c/Joaquin Costa 22, 28002 Madrid, Spain
Mr M.A. DEL SAZ DE LA FUENTE	Picos de Europa 7, Poligono Industrial de S. Fernando de Henares II, 28850 Torrejon de Ardoz, Madrid, Spain
Ms D. DELEUZE*	Construcciones Aeronauticas S.A. (CASA), R&D Technology and Materials Dept., Getafe 28065 Madrid, Spain
Mr G. DESROCQUES†	LIFO Universite d'Orléans, B.P. 6759, 45067 Orléans Cedex 2, France
Mr G. DI MARTINO*	CIRA, Relazioni Esterne, Via Boncompagni 93, 00187 Rome, Italy
Dr C.M. DIAS	Universidad da Beira Interior, Rua Marques D'Avila e Bolama, 6200 Covilha, Portugal
Mrs M. ECHANOVE	Jefe Centro Documentacion, Fundacion Mapfre Estudios, Monte del Pilar s/n, 28023 Madrid, Spain
Mr S.D. ELPHICK	Head of Library, Royal Netherlands Naval College, Het Nieuwe Diep 8, 1781 AC Den Helder, The Netherlands
Captain J.M. EVANS	Office of Naval Research, European Office, 223/231 Old Marylebone Road, London NW1 5TH, United Kingdom
Ms G. FAINSTEIN	ISOC, c/Pinar 25, 28006 Madrid, Spain
Mr G. FARQUHAR†	Defence Research Information Centre, Room 2164, Kentigern House, 65 Brown Street, Glasgow G2 8EX, United Kingdom
Mrs C. FERNANDEZ GALIANO	Universidad Alcala de Henares Biblioteca, Edificio de Medicina, Km 33 — N.II, 28871 Alcala de Henares, Spain
Ms C. FERRER	Head of Computer S. Dept., Plaza Honduras 6, 18, 46022 Valencia, Spain
Prof. C. FLUHR†	INSTN/MIST, CEN/Saclay, 91119 Gif-sur-Yvette, Cedex, France
Ms C. GALBAN	Documentalista, I.C.Y.T., c/Joaquin Costa 22, 28002 Madrid, Spain
Ms A. GALLEGO GARCIA	CASA, c/Princesa 47-1, Madrid, Spain
Mr M.A. GARCIA ALONSO	Projects Manager, CEMI, Conde Duque 9, 28015 Madrid, Spain
Mr J.J. GEORGET	Aérospatiale (DSSS), MU/DTD Service Documentation, B.P. 96, 78133 Les Mureaux Cedex, France
Ms M.J. GOMEZ SANCHEZ	Centro de Calculo, Facultad de Ciencias, Ctra. Madrid-Barcelona Km 33, 600, 28871 Alcala de Henares (Madrid), Spain
Ms M. GONZALEZ GONZALEZ	Comision Nacional del Mercado de Valores, Pa. de la Castellana 19, 28046 Madrid, Spain
Mr W.F.M. GREGOIRE	Chief Central Records, Seccos Secretariat, B-7010 SHAPE, Belgium
Dr M.C. GUTIERREZ*	INTA, Paseo Pintor Rosales 34, 28008 Madrid, Spain
Ms H. HALLER†	Defense Technical Information Center — Defense Logistics Agency, Cameron Station, Alexandria VA 22304-6145, United States
Mr T.A. HERMANN*	Head, Customer Service & Marketing, NASA Center for Aerospace Information, P.O. Box 8757, BWI Airport, MD 21240, United States
Mr J.C. HERNANDEZ ALVAREZ	Information Scientist, C/Canillas 70-1-D, 28002 Madrid, Spain
Mrs R. HERRERO PEREZ	Directora, Mo. Industria, Comercia y Turismo Biblioteca General, Po. Castellana 160, 28071 Madrid, Spain
Ms A. HERVAS	c/Fco. Silvela 90 4-B, 28062 Madrid, Spain
Mr B. HISINGER, M.Sc.E.E.*	Senior Consultant/The National Technological Library Denmark, c/o Technical University of Denmark, Anker Engelunds Vej 1, 2800 Lyngby, Denmark
Ir R.A. HOGENDOORN†	Dept. IN, National Aerospace Laboratory (NLR), P.O. Box 153, 8300 AD Emmeloord, The Netherlands
Ir P.J. HOOGENBERK*	Ministerie van Defensie, Hoofd TDCK, Postbox 90701, 2509 LS The Hague, The Netherlands

† Author
* Panel Member

Mrs A. HUESCA GONZALEZ	I.E.S.A., Alfonso XII, 18, 5, 28014 Madrid, Spain
Dr J.M. HUGHES*	Head, Technical Library, Code E23, Naval Surface Warfare Center, Dehlgren VA 22448-5020, United States
Mrs D. JULE	Chef du Centre Documentation Aérospatiale, 12 rue Beranger, 92322 Châtillon, France
Mrs S. JULIAN	Plan Nacional de I&D, c/Rosario Pino 14—16, 28020 Madrid, Spain
Ms C. JULIAN	SEDIC, Gran Via 67, 28028 Madrid, Spain
Ms K. KAYE†	NASA Scientific and Technical Information Program, Code NTT, Washington D.C. 20546, United States
Mrs A. KNECHT CRISTOBAL	Instituto Juan March de Estudios e Investigaciones, Spain
Prof. Dr G.E. KNORZ†	Fachhochschule Darmstadt, Department of Computer Science IuD, Schofferstrasse 3, W-6100 Darmstadt, Germany
Mr A. KUHN†	NASA Scientific and Technical Information Program, Code NTT, Washington D.C. 20546, United States
Mr O. LAVROFF	5 rue Anna Jacquin, 92100 Boulogne, France
Mrs C. LAZARO CORTHAY	Centro de Documentacion, Ministerio de Defensa, Plaza de la Castellana 109, 28046 Madrid, Spain
Mr P. LAZARO DELGADO	CASA, Proyectos I-DTM, Getafe 28065 Madrid, Spain
Mrs F. LHULLIER*	Chef du Service Documentation, ONERA, 29 Av. de la Division Leclerc, 92320 Châtillon, France
Mrs M.L. LOPEZ MARTINEZ	Jefe Seccion Biblioteca, Po. Prado 6, 28014 Madrid, Spain
Mr E. LOPEZ DE QUINTANA	Jefe del Centro de Documentacion, Ctz. Madrid-Irun, Km 19.3, Spain
Ms C. MACHARD†	INSTN/MIST, CEN/Saclay, 91119 Gif-sur-Yvette, Cedex, France
Ms M.E. MARTIN SANZ MARTINEZ	Delegada ESA-IRS, INTA, 34 Paseo Pintor Rosales, 28008 Madrid, Spain
Ms R. MARTIN Y DE VEGA	Centro de Documentacion, Ministerio de Defensa, Po. de la Castellana 109, 28071 Madrid, Spain
Mr M.A. MARTINEZ	Director General, Fundacion Mapfre Estudios s/n, Monte del Pilar 5, 28023 El Plantio, Spain
Mr J.M. MARTINEZ JUDEZ	CASA — Proyectos I-DTM, Getafe 28065 Madrid, Spain
Ms M.C. MAZAUD**	70 rue de Paris, 92100 Boulogne, France
Ms E. MAZORRA DE QUERO	General Lacy 22-2-A, 28045 Madrid, Spain
Mr A. MEKAOUCHE†	LIFO Université d'Orléans, B.P. 6759, 45067 Orléans Cedex 2, France
Mr J. MOLINA BENAYAS	I.E.S.A., Alfonso XII, 18, 5, 28014 Madrid, Spain
Ms S. MONREAL REQUENA	INTA, Paseo Pintor Rosales 34, 28008 Madrid, Spain
Ms A. OZDIL*	Ministry of National Defence (MSB) Department of R&D (ARGE), 06550 Ankara, Turkey
Mrs F. PALLARDO	Jefa Seccion de Biblioteca, Secretaria General de Medio Ambiente, Po. Castellana 67, 28071 Madrid, Spain
Mrs D. PATRINOU*	Technical Scientific Library, Hellenic Aerospace Industry, P.O. Box 23, GR 32009 Schimatari, Viotias, Greece
Ms A. PEREZ MATO	Biblioteca Central de Experimentale, Universidad de Alcala de Henares, Km 33 de N II, 28871 Alcala de Henares, Spain
Dr L. PEZZI	Centro Nacional de Biotechnologia, Campus Universidad Autonoma, 28049 Madrid, Spain
Mrs E. PRIMO PENA	C.N. Farmacobiologia, Instituto de Salud Carlos III, 28220 Majadahonda (Madrid), Spain
Lt Col. J.G. RENY*	NATO Terminology Coordinator, NATO/MAS, B-1110 Brussels, Belgium

† Author
* Panel Member
** Interpreter

Prof. C.J. VAN RIJSBERGEN†	Department of Computing Science, University of Glasgow, 8—17 Lilybank Gardens, Glasgow G12 8RZ, United Kingdom
Ms M. RIVAS VARA	Documentalista, Centro de Documentacion, Ministerio de Defensa, Po. de la Castellana 109, 28071 Madrid, Spain
Mr R. RODRIGUEZ DE CORA	Consultant, Calle Infanta Maria Teresa 6-5-0, 28016 Madrid, Spain
Mr C.J. RODRIGUEZ	c/Fernando el Catolico 10-3D, 28015 Madrid, Spain
Mrs C. ROMEO PEREZ	Documentalista, Centro Nacional Farmacobiologia, Instituto Salud Carlos III, Majadahonda, 28220 Madrid, Spain
Mrs E. RUBIO	Apartado 294, 28850 Torrejon Ardoz, Spain
Mr M. SAINZ DE LOS TERREROS	Servicio de Banco de Datos de Medicamentos, Pl 9a, Ministerio de Sanidad y Consumo, Paseo del Prado 18—20, 28071 Madrid, Spain
Mr C. SALMON	Chef du Service Documentation, Laborelec, 125 rue de Rhode, B-1630 Linkerbeek, Belgium
Ms R. SANNINO	C.I.R.A., Via Maiorise, 81043 Capua (CE), Italy
Mrs M.C. SANZ BOMBIN	Secretaria Tecnica de Transportes y Comunicaciones — Biblioteca, Ministerio de Obras Publicas, Plaza de San Juan de la Cruz s/n, 28003 Madrid, Spain
Mr M.J. SCHRYER*	Director, Directorate of Scientific Information Services, National Defence Headquarters, MGeneral George R. Pearkes Building, Ottawa, Ontario K1A 0K2, Canada
Mrs D. SCHRYER	41 Pheasant Run Drive, Nepean, Ontario K2J 2R3, Canada
Mr R. SEARLE (Deputy Panel Chairman)	Chief Librarian, Royal Aerospace Establishment, Farnborough, Hants GU14 6TD, United Kingdom
Mr D. SHEARER**	3 Impasse Robert, 75018 Paris, France
Ms A. SISNIEGA	Centro de Documentacion del INI, Plaza de Salamanca 8, 28006 Madrid, Spain
Mr R.W. SLANEY	Head of Library, Defence Operational Analysis Est., Parvis Road, West Byfleet, Surrey KT14 6LY, United Kingdom
Mr A.W. SMITH	The British Library DSC, Boston Spa, Wetherby, West Yorkshire LS23 7BQ, United Kingdom
Dr L. SOTO BACA	Documentacion Medica, c/Almirante Cadarso 8—3a, 46005 Valencia, Spain
Dr M.P. SOUSA	Universidade da Beira Interior, Rua Marques D'Avila e Bolama, 6200 Covilha, Portugal
Mrs M. DE SUSBIELLE**	11 quai Paul Doumer, 92400 Courbevoie, France
Ir A.S.T. TAN*	Information Specialist, National Aerospace Laboratory (NLR), P.O. Box 90502, 1006 BM Amsterdam, The Netherlands
Prof. I. TENORIO VAZQUEZ	Facultad de Ciencias de la Informacion, c/Gonzalo Bilbao 7 y 9, 41003 Sevilla, Spain
Ms H. TUIT	TNO-PML, P.O. Box 45, 2200 AA Ryswizk, The Netherlands
Dr N. TUNCER*	Yök Dokümantasyon Merkezi, 06530-Bilkent, Ankara, Turkey
Mr L. URQUIZA ECHEVARREN	Director General, Mas Menos Uno, S.L., Natalia de Silva 3, 28027 Madrid, Spain
Mr J. VALENCIA	Aérospatiale, B.P.2, 78133 Les Mureaux Cedex, France
Ms C.M. VARELA SOUTO	INTA, Paseo Pintor Rosales 34, 28008 Madrid, Spain
Mrs I. VENAKI-MELINTZI*	Hellenic Air Force General Staff, Branch B'/Directorate B5/Section 1, Messogion Avenue, Holargos TGA 1010, Athens, Greece
Mr G.C. VIS	FEL-TNO, Postbus 96864, 2509 JG The Hague, The Netherlands
Ms C. WALKER*	Head, Information Services Branch, SHAPE Technical Centre, P.O. Box 174, 2501 CD The Hague, The Netherlands

† Author
* Panel Member
** Interpreter

Dr A.J. WENNERSTROM	Director, AGARD, 7 rue Ancelle, 92200 Neuilly-sur-Seine, France
Mrs C. WIEGANDT*	Société Aérospatiale, 12 rue Pasteur — BP 76, 92152 Suresnes Cedex, France
Dr U. WIEMKEN	INT, Appelsgarten 2, D-5350 Euskirchen, Germany
Mr M.R.C. WILKINSON*	Head, Defence Research Information Centre, Kentigern House, 65 Brown Street, Glasgow G2 8EX, United Kingdom
Mr G.H. WOOD†	Manager, Scientific Numeric D/B Service, Canada Institute for Scientific and Technical Information, National Research Council of Canada, Ottawa, Ontario K1A 0S2, Canada
Mr A. YANEZ (Panel Chairman)	Conseiller du Directeur, C.E.D.O.C.A.R., 00460 Armées, France
Ir B.H.A. ZIJLSTRA†	Scientific/Technical Documentation Centre of the Netherlands Armed Forces (TDCK), P.O. Box 90701, 2509 LS The Hague, The Netherlands

† Author
*** Panel Member**

REPORT DOCUMENTATION PAGE

1. Recipient's Reference	2. Originator's Reference	3. Further Reference	4. Security Classification of Document
	AGARD-CP-505	ISBN 92-835-0655-3	UNCLASSIFIED

5. Originator Advisory Group for Aerospace Research and Development
North Atlantic Treaty Organization
7 rue Ancelle, 92200 Neuilly sur Seine, France

6. Title BRINGING DOWN THE BARRIERS TO INFORMATION TRANSFER

7. Presented at the Technical Information Panel Specialists' Meeting held at the Instituto Nacional de Industria, Madrid, Spain 8th—9th October 1991.

8. Author(s)/Editor(s)	9. Date
Various	February 1992

10. Author's/Editor's Address	11. Pages
Various	146

12. Distribution Statement This document is distributed in accordance with AGARD policies and regulations, which are outlined on the back covers of all AGARD publications.

13. Keywords/Descriptors

Information technology
Information retrieval
Information systems
Expert systems
Optical character recognition device
Data compression
Natural language
Computer programs
Gateways — telecommunication

14. Abstract

Comprises the papers presented at a Specialists' Meeting held by the Technical Information Panel of AGARD in Madrid, 8th to 9th October 1991.

Topics discussed include intelligent gateways, expert systems, new methods of information retrieval, including non-boolean and natural language approaches and hypertext, optical character recognition, data compression, and computer systems for handling engineering property data. There was also a keynote address serving as an overview of these topics, and the proceedings contain a technical evaluation report.

AGARD Conference Proceedings 505
Advisory Group for Aerospace Research and Development, NATO
BRINGING DOWN THE BARRIERS TO INFORMATION TRANSFER
Published February 1992
146 pages

Comprises the papers presented at a Specialists' Meeting held by the Technical Information Panel of AGARD in Madrid, 8th to 9th October 1991.

Topics discussed include intelligent gateways, expert systems, new methods of information retrieval, including non-boolean and natural language approaches and hypertext, optical character recognition, data

P.T.O.

AGARD-CP-505

Information technology
Information retrieval
Information systems
Expert systems
Optical character recognition device
Data compression
Natural language
Computer programs
Gateways — telecommunication

AGARD Conference Proceedings 505
Advisory Group for Aerospace Research and Development, NATO
BRINGING DOWN THE BARRIERS TO INFORMATION TRANSFER
Published February 1992
146 pages

Comprises the papers presented at a Specialists' Meeting held by the Technical Information Panel of AGARD in Madrid, 8th to 9th October 1991.

Topics discussed include intelligent gateways, expert systems, new methods of information retrieval, including non-boolean and natural language approaches and hypertext, optical character recognition, data

P.T.O.

AGARD-CP-505

Information technology
Information retrieval
Information systems
Expert systems
Optical character recognition device
Data compression
Natural language
Computer programs
Gateways — telecommunication

AGARD Conference Proceedings 505
Advisory Group for Aerospace Research and Development, NATO
BRINGING DOWN THE BARRIERS TO INFORMATION TRANSFER
Published February 1992
146 pages

Comprises the papers presented at a Specialists' Meeting held by the Technical Information Panel of AGARD in Madrid, 8th to 9th October 1991.

Topics discussed include intelligent gateways, expert systems, new methods of information retrieval, including non-boolean and natural language approaches and hypertext, optical character recognition, data

P.T.O.

AGARD-CP-505

Information technology
Information retrieval
Information systems
Expert systems
Optical character recognition device
Data compression
Natural language
Computer programs
Gateways — telecommunication

AGARD Conference Proceedings 505
Advisory Group for Aerospace Research and Development, NATO
BRINGING DOWN THE BARRIERS TO INFORMATION TRANSFER
Published February 1992
146 pages

Comprises the papers presented at a Specialists' Meeting held by the Technical Information Panel of AGARD in Madrid, 8th to 9th October 1991.

Topics discussed include intelligent gateways, expert systems, new methods of information retrieval, including non-boolean and natural language approaches and hypertext, optical character recognition, data

P.T.O.

AGARD-CP-505

Information technology
Information retrieval
Information systems
Expert systems
Optical character recognition device
Data compression
Natural language
Computer programs
Gateways — telecommunication

compression, and computer systems for handling engineering property data. There was also a keynote address serving as an overview of these topics, and the proceedings contain a technical evaluation report.

ISBN 92-835-0655-3

compression, and computer systems for handling engineering property data. There was also a keynote address serving as an overview of these topics, and the proceedings contain a technical evaluation report.

ISBN 92-835-0655-3

compression, and computer systems for handling engineering property data. There was also a keynote address serving as an overview of these topics, and the proceedings contain a technical evaluation report.

ISBN 92-835-0655-3

compression, and computer systems for handling engineering property data. There was also a keynote address serving as an overview of these topics, and the proceedings contain a technical evaluation report.

ISBN 92-835-0655-3

AGARD

NATO · OTAN

7 RUE ANCELLE · 92200 NEUILLY-SUR-SEINE

FRANCE

Téléphone (1)47.38.57.00 · Télex 610 176
Télécopie (1)47.38.57.99

DIFFUSION DES PUBLICATIONS AGARD NON CLASSIFIEES

L'AGARD ne détient pas de stocks de ses publications, dans un but de distribution générale à l'adresse ci-dessus. La diffusion initiale des publications de l'AGARD est effectuée auprès des pays membres de cette organisation par l'intermédiaire des Centres Nationaux de Distribution suivants. A l'exception des Etats-Unis, ces centres disposent parfois d'exemplaires additionnels; dans les cas contraire, on peut se procurer ces exemplaires sous forme de microfiches ou de microcopies auprès des Agences de Vente dont la liste suite.

CENTRES DE DIFFUSION NATIONAUX

ALLEMAGNE
Fachinformationszentrum,
Karlsruhe
D-7514 Eggenstein-Leopoldshafen 2

BELGIQUE
Coordonnateur AGARD-VSL
Etat-Major de la Force Aérienne
Quartier Reine Elisabeth
Rue d'Evere, 1140 Bruxelles

CANADA
Directeur du Service des Renseignements Scientifiques
Ministère de la Défense Nationale
Ottawa, Ontario K1A 0K2

DANEMARK
Danish Defence Research Board
Ved Idraetsparken 4
2100 Copenhagen Ø

ESPAGNE
INTA (AGARD Publications)
Pintor Rosales 34
28008 Madrid

ETATS-UNIS
National Aeronautics and Space Administration
Langley Research Center
M/S 180
Hampton, Virginia 23665

FRANCE
O.N.E.R.A. (Direction)
29, Avenue de la Division Leclerc
92320, Châtillon sous Bagneux

GRECE
Hellenic Air Force
Air War College
Scientific and Technical Library
Dekelia Air Force Base
Dekelia, Athens TGA 1010

ISLANDE
Director of Aviation
c/o Flugrad
Reykjavik

ITALIE
Aeronautica Militare
Ufficio del Delegato Nazionale all'AGARD
Aeroporto Pratica di Mare
00040 Pomezia (Roma)

LUXEMBOURG
Voir Belgique

NORVEGE
Norwegian Defence Research Establishment
Attn: Biblioteket
P.O. Box 25
N-2007 Kjeller

PAYS-BAS
Netherlands Delegation to AGARD
National Aerospace Laboratory NLR
Kluyverweg 1
2629 HS Delft

PORTUGAL
Portuguese National Coordinator to AGARD
Gabinete de Estudos e Programas
CLAFA
Base de Alfragide
Alfragide
2700 Amadora

ROYAUME UNI
Defence Research Information Centre
Kentigern House
65 Brown Street
Glasgow G2 8EX

TURQUIE
Milli Savunma Başkanlığı (MSB)
ARGE Daire Başkanlığı (ARGE)
Ankara

LE CENTRE NATIONAL DE DISTRIBUTION DES ETATS-UNIS (NASA) NE DETIENT PAS DE STOCKS DES PUBLICATIONS AGARD ET LES DEMANDES D'EXEMPLAIRES DOIVENT ETRE ADRESSEES DIRECTEMENT AU SERVICE NATIONAL TECHNIQUE DE L'INFORMATION (NTIS) DONT L'ADRESSE SUIT.

AGENCES DE VENTE

National Technical Information Service
(NTIS)
5285 Port Royal Road
Springfield, Virginia 22161
Etats-Unis

ESA/Information Retrieval Service
European Space Agency
10, rue Mario Nikis
75015 Paris
France

The British Library
Document Supply Division
Boston Spa, Wetherby
West Yorkshire LS23 7BQ
Royaume Uni

Les demandes de microfiches ou de photocopies de documents AGARD (y compris les demandes faites auprès du NTIS) doivent comporter la dénomination AGARD, ainsi que le numéro de série de l'AGARD (par exemple AGARD-AG-315). Des informations analogues, telles que le titre et la date de publication sont souhaitables. Veuiller noter qu'il y a lieu de spécifier AGARD-R-nnn et AGARD-AR-nnn lors de la commande de rapports AGARD et des rapports consultatifs AGARD respectivement. Des références bibliographiques complètes ainsi que des résumés des publications AGARD figurent dans les journaux suivants:

Scientifique and Technical Aerospace Reports (STAR)
publié par la NASA Scientific and Technical
Information Division
NASA Headquarters (NTT)
Washington D.C. 20546
Etats-Unis

Government Reports Announcements and Index (GRA&I)
publié par le National Technical Information Service
Springfield
Virginia 22161
Etats-Unis
(accessible également en mode interactif dans la base de données bibliographiques en ligne du NTIS, et sur CD-ROM)

Imprimé par Specialised Printing Services Limited
40 Chigwell Lane, Loughton, Essex IG10 3TZ

NATO OTAN
7 RUE ANCELLE · 92200 NEUILLY-SUR-SEINE
FRANCE

Telephone (1)47.38.57.00 · Telex 610 176
Telefax (1)47.38.57.99

DISTRIBUTION OF UNCLASSIFIED AGARD PUBLICATIONS

AGARD does NOT hold stocks of AGARD publications at the above address for general distribution. Initial distribution of AGARD publications is made to AGARD Member Nations through the following National Distribution Centres. Further copies are sometimes available from these Centres (except in the United States), but if not may be purchased in Microfiche or Photocopy form from the Sales Agencies listed below.

NATIONAL DISTRIBUTION CENTRES

BELGIUM
Coordonnateur AGARD — VSL
Etat-Major de la Force Aérienne
Quartier Reine Elisabeth
Rue d'Evere, 1140 Bruxelles

CANADA
Director Scientific Information Services
Dept of National Defence
Ottawa, Ontario K1A 0K2

DENMARK
Danish Defence Research Board
Ved Idraetsparken 4
2100 Copenhagen Ø

FRANCE
O.N.E.R.A. (Direction)
29 Avenue de la Division Leclerc
92320 Châtillon

GERMANY
Fachinformationszentrum
Karlsruhe
D-7514 Eggenstein-Leopoldshafen 2

GREECE
Hellenic Air Force
Air War College
Scientific and Technical Library
Dekelia Air Force Base
Dekelia, Athens TGA 1010

ICELAND
Director of Aviation
c/o Flugrad
Reykjavik

ITALY
Aeronautica Militare
Ufficio del Delegato Nazionale all'AGARD
Aeroporto Pratica di Mare
00040 Pomezia (Roma)

LUXEMBOURG
See Belgium

NETHERLANDS
Netherlands Delegation to AGARD
National Aerospace Laboratory, NLR
Kluyverweg 1
2629 HS Delft

NORWAY
Norwegian Defence Research Establishment
Attn: Biblioteket
P.O. Box 25
N-2007 Kjeller

PORTUGAL
Portuguese National Coordinator to AGARD
Gabinete de Estudos e Programas
CLAFA
Base de Alfragide
Alfragide
2700 Amadora

SPAIN
INTA (AGARD Publications)
Pintor Rosales 34
28008 Madrid

TURKEY
Milli Savunma Başkanlığı (MSB)
ARGE Daire Başkanlığı (ARGE)
Ankara

UNITED KINGDOM
Defence Research Information Centre
Kentigern House
65 Brown Street
Glasgow G2 8EX

UNITED STATES
National Aeronautics and Space Administration (NASA)
Langley Research Center
M/S 180
Hampton, Virginia 23665

THE UNITED STATES NATIONAL DISTRIBUTION CENTRE (NASA) DOES NOT HOLD STOCKS OF AGARD PUBLICATIONS, AND APPLICATIONS FOR COPIES SHOULD BE MADE DIRECT TO THE NATIONAL TECHNICAL INFORMATION SERVICE (NTIS) AT THE ADDRESS BELOW.

SALES AGENCIES

National Technical
Information Service (NTIS)
5285 Port Royal Road
Springfield, Virginia 22161
United States

ESA/Information Retrieval Service
European Space Agency
10, rue Mario Nikis
75015 Paris
France

The British Library
Document Supply Centre
Boston Spa, Wetherby
West Yorkshire LS23 7BQ
United Kingdom

Requests for microfiches or photocopies of AGARD documents (including requests to NTIS) should include the word 'AGARD' and the AGARD serial number (for example AGARD-AG-315). Collateral information such as title and publication date is desirable. Note that AGARD Reports and Advisory Reports should be specified as AGARD-R-nnn and AGARD-AR-nnn, respectively. Full bibliographical references and abstracts of AGARD publications are given in the following journals:

Scientific and Technical Aerospace Reports (STAR)
published by NASA Scientific and Technical
Information Division
NASA Headquarters (NTT)
Washington D.C. 20546
United States

Government Reports Announcements and Index (GRA&I)
published by the National Technical Information Service
Springfield
Virginia 22161
United States
(also available online in the NTIS Bibliographic Database or on CD-ROM)

Printed by Specialised Printing Services Limited
40 Chigwell Lane, Loughton, Essex IG10 3TZ

ISBN 92-835-0655-3